

REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY) 31-08-2015	2. REPORT TYPE Ph.D. Dissertation	3. DATES COVERED (From - To) -		
4. TITLE AND SUBTITLE NETWORK CHARACTERISTICS AND DYNAMICS: RECIPROCITY, COMPETITION AND INFORMATION DISSEMINATION		5a. CONTRACT NUMBER W911NF-12-1-0385		
		5b. GRANT NUMBER		
		5c. PROGRAM ELEMENT NUMBER 611103		
6. AUTHORS B. Jiang		5d. PROJECT NUMBER		
		5e. TASK NUMBER		
		5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAMES AND ADDRESSES Cornell University Office of Sponsored Programs 373 Pine Tree Road Ithaca, NY 14850 -2820		8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211		10. SPONSOR/MONITOR'S ACRONYM(S) ARO		
		11. SPONSOR/MONITOR'S REPORT NUMBER(S) 61783-MA-MUR.104		
12. DISTRIBUTION AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.				
13. SUPPLEMENTARY NOTES The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.				
14. ABSTRACT Networks are commonly used to study complex systems. This often requires a good understanding of the structural characteristics and evolution dynamics of networks, and also their impacts on a variety of dynamic processes taking place on top of them. In this thesis, we study various aspects of network characteristics and dynamics, with a focus on reciprocity, competition and information dissemination. We first formulate the maximum reciprocity problem and study its use in the interpretation of reciprocity in real networks. We propose to interpret reciprocity based on its comparison with the maximum possible reciprocity for a				
15. SUBJECT TERMS network dynamics, competition, reciprocity, information dissemination				
16. SECURITY CLASSIFICATION OF: a. REPORT UU		17. LIMITATION OF ABSTRACT UU	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Sidney Resnick
b. ABSTRACT UU		c. THIS PAGE UU		19b. TELEPHONE NUMBER 607-255-1210

Report Title

NETWORK CHARACTERISTICS AND DYNAMICS: RECIPROCITY, COMPETITION AND INFORMATION DISSEMINATION

ABSTRACT

Networks are commonly used to study complex systems. This often requires a good understanding of the structural characteristics and evolution dynamics of networks, and also their impacts on a variety of dynamic processes taking place on top of them. In this thesis, we study various aspects of network characteristics and dynamics, with a focus on reciprocity, competition and information dissemination.

We first formulate the maximum reciprocity problem and study its use in the interpretation of reciprocity in real networks. We propose to interpret reciprocity based on its comparison with the maximum possible reciprocity for a network exhibiting the same degrees. We show that the maximum reciprocity problem is NP-hard, and use an upper bound instead of the maximum. We find that this bound is surprisingly close to the empirical reciprocity in a wide range of real networks, and that there is

via surprisingly strong linear relationship between the two. We also show that certain small suboptimal motifs called 3-paths are the major cause for suboptimality in real networks.

Secondly, we analyze competition dynamics under cumulative advantage, where accumulated resource promotes gathering even more resource. We characterize the tail distributions of duration and intensity for pairwise competition. We show that duration always has a power-law tail irrespective of competitors' fitness, while intensity has either a power-law tail or an exponential tail depending on whether the competitors are equally fit. We observe a struggle-of-the-fitness phenomenon, where a slight difference in fitness results in an extremely heavy tail of duration distribution. Lastly, we study the efficiency of information dissemination in social networks with limited budget of attention. We quantify the efficiency of information dissemination for both cooperative and selfish user behaviors in various network topologies. We identify topologies where cooperation plays a critical role in efficient information propagation. We propose an incentive mechanism called "plus-one" to coax users into cooperation in such cases.

**NETWORK CHARACTERISTICS AND DYNAMICS:
RECIPROCITY, COMPETITION AND INFORMATION
DISSEMINATION**

A Dissertation Presented

by

BO JIANG

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2015

College of Information and Computer Sciences

© Copyright by Bo Jiang 2015

All Rights Reserved

**NETWORK CHARACTERISTICS AND DYNAMICS:
RECIPROCITY, COMPETITION AND INFORMATION
DISSEMINATION**

A Dissertation Presented

by

BO JIANG

Approved as to style and content by:

Don Towsley, Chair

Weibo Gong, Member

Matthias Grossglauser, Member

David Jensen, Member

Ramesh Sitaraman, Member

James Allan, Chair of the Faculty
College of Information and Computer Sciences

To my parents.

ACKNOWLEDGMENTS

I would like to express my deepest thanks to my advisor, Prof. Don Towsley, who has always been encouraging, supportive and inspirational throughout the years of my PhD study. I thank Prof. Weibo Gong for numerous enjoyable and mind-opening discussions. I thank Prof. Matthias Grossglauser, Prof. David Jensen, Prof. Ramesh Sitaraman, for serving on my committee and for their valuable comments and suggestions. I thank Prof. Neil Immerman and Prof. Dennis Goeckel for being my portfolio letter writers.

I am grateful to Dr. Laurent Massoulié and Dr. Nidhi Hegde for being my mentors during my internship at Technicolor Paris Research Lab. I thank Prof. Zhi-Li Zhang, Prof. Daniel Figueiredo, Dr. Bruno Ribeiro, Liyuan Sun, for stimulating discussions and fruitful collaborations on work presented in this thesis.

I thank all my colleagues in the Networks Lab and all my friends at UMass, who have made my time at UMass an unforgettable experience. My dearest friends outside UMass, Han Wang, Yihong Wu, Yu Zhang, Shuo Guo, are always there for me during both joyful and stressful times, to whom I am always thankful.

Last but not least, I would like to express my deepest gratitude to my parents and all my extended family members for their constant love and support.

This work was supported in part by DoD ARO MURI Award W911NF-12-1-0385 and NSF grant CNS-1065133.

ABSTRACT

NETWORK CHARACTERISTICS AND DYNAMICS: RECIPROCITY, COMPETITION AND INFORMATION DISSEMINATION

SEPTEMBER 2015

BO JIANG

B.Sc., TSINGHUA UNIVERSITY, BEIJING, CHINA

M.Sc., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Don Towsley

Networks are commonly used to study complex systems. This often requires a good understanding of the structural characteristics and evolution dynamics of networks, and also their impacts on a variety of dynamic processes taking place on top of them. In this thesis, we study various aspects of network characteristics and dynamics, with a focus on reciprocity, competition and information dissemination.

We first formulate the maximum reciprocity problem and study its use in the interpretation of reciprocity in real networks. We propose to interpret reciprocity based on its comparison with the maximum possible reciprocity for a network exhibiting the same degrees. We show that the maximum reciprocity problem is NP-hard, and use an upper bound instead of the maximum. We find that this bound is surprisingly close to the empirical reciprocity in a wide range of real networks, and that there is

a surprisingly strong linear relationship between the two. We also show that certain small suboptimal motifs called 3-paths are the major cause for suboptimality in real networks.

Secondly, we analyze competition dynamics under cumulative advantage, where accumulated resource promotes gathering even more resource. We characterize the tail distributions of duration and intensity for pairwise competition. We show that duration always has a power-law tail irrespective of competitors' fitness, while intensity has either a power-law tail or an exponential tail depending on whether the competitors are equally fit. We observe a struggle-of-the-fitness phenomenon, where a slight difference in fitness results in an extremely heavy tail of duration distribution.

Lastly, we study the efficiency of information dissemination in social networks with limited budget of attention. We quantify the efficiency of information dissemination for both cooperative and selfish user behaviors in various network topologies. We identify topologies where cooperation plays a critical role in efficient information propagation. We propose an incentive mechanism called “plus-one” to coax users into cooperation in such cases.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	v
ABSTRACT	vi
LIST OF TABLES	xi
LIST OF FIGURES.....	xii
 CHAPTER	
1. INTRODUCTION	1
1.1 Thesis Contributions	3
2. RECIPROCITY IN NETWORKS WITH DEGREE CONSTRAINTS	5
2.1 Introduction	5
2.2 Graphic Sequences and Maximum Reciprocity Problem	8
2.2.1 Graphic Degree and Bi-degree Sequences	8
2.2.2 Maximum Reciprocity Problem	9
2.2.3 Some Notations	10
2.3 Hardness Analysis and Bounds	11
2.3.1 Upper Bound for Reciprocity	12
2.3.2 Proof of NP-hardness	14
2.3.3 Sufficient Conditions for Achieving Upper Bound	17
2.4 Patterns in Maximum Digraphs	18
2.4.1 Small Suboptimal Motifs	19
2.4.2 Properties of Maximum Digraphs.....	21
2.4.2.1 3-path Optimal Digraphs.....	21

2.4.2.2	Maximum Digraphs	23
2.4.3	Some Examples	26
2.5	Empirical Study	29
2.5.1	Datasets	29
2.5.2	Empirical Reciprocity vs. Upper Bound	30
2.5.3	Reciprocity of 3-path Optimal Digraphs	33
2.6	Conclusion	35
3.	COMPETITION UNDER CUMULATIVE ADVANTAGE	36
3.1	Introduction	36
3.2	Related Work	39
3.3	Models	40
3.3.1	General Setup and Metrics	40
3.3.2	CA Competition Model	42
3.4	Results	44
3.4.1	Competition Duration	45
3.4.1.1	Equal Fitness Case: CA ₌	45
3.4.1.2	Different Fitness Case: CA _≠	47
3.4.1.3	Struggle-of-the-Fittest Phenomenon	49
3.4.2	Competition Intensity	50
3.4.2.1	Equal Fitness Case: CA ₌	50
3.4.2.2	Different Fitness Case: CA _≠	51
3.4.3	Interplay of Duration and Intensity	53
3.5	Discussion and Conclusion	55
4.	INFORMATION DISSEMINATION IN SOCIAL NETWORKS UNDER LIMITED BUDGET OF ATTENTION	58
4.1	Introduction	58
4.2	Model Description	60
4.3	Efficiency Analysis	63
4.3.1	Tree Network	63

4.3.1.1	Line Network	65
4.3.1.2	Chained Star Network	66
4.3.1.3	k -ary Tree Network.....	66
4.3.2	Clique Networks	67
4.3.3	Expander Network	71
4.4	Incentivizing Efficient Behavior	74
4.4.1	The Plus-One mechanism	74
4.4.2	Inefficient Suboptimal Graphs.....	77
4.5	Simulation Results	77
4.6	Conclusion and Future Work	80
5.	CONCLUSIONS AND FUTURE WORK	82
 APPENDICES		
A.	ADDITIONAL PROOFS FOR CHAPTER 2	84
B.	DATASETS IN SECTION 2.5	96
C.	ADDITIONAL PROOFS FOR CHAPTER 3	100
D.	ADDITIONAL PROOFS FOR CHAPTER 4	115
 BIBLIOGRAPHY		122

LIST OF TABLES

Table	Page
3.1 Tail distributions for duration and intensity of competitions in both RW and CA models. Multiplicative constants are omitted in all expressions involving t and n . The RW statistics can be found in most textbooks on the topic, e.g. [38, pp. 113,116].	45
B.1 Statistics of some real networks.	96

LIST OF FIGURES

Figure	Page
2.1 Graphic bi-sequence with non-graphic max and min sequences.	13
2.2 Graphic bi-sequence with graphic min sequence but non-graphic max sequence.	13
2.3 Graphic bi-sequence with graphic max sequence but non-graphic min sequence.	14
2.4 Insufficiency of necessary condition in Proposition 2.3.1.	15
2.5 Different types of 3-paths with corresponding rewirings.	19
2.6 Suboptimal even cycle with reciprocated edges.	24
2.7 Patterns of even paths in maximum digraphs.	26
2.8 Example 2.4.11.	28
2.9 Example 2.4.12.	29
2.10 Scatter plot of empirical reciprocity versus upper bound.	30
2.11 Scatter plot of number of reciprocated edges versus upper bound.	31
2.12 Box plot of reciprocity-bound ratio for different network categories.	32
2.13 Scatter plot of reciprocity of the 3-path optimal digraph returned by Algorithm 1 versus upper bound.	34
2.14 Box plot of ratio between reciprocity of 3-path optimal digraph returned by Algorithm 1 and upper bound.	34
3.1 Time evolution of degree difference between two nodes in competitions without and with cumulative advantage.	38

3.2	State space of competition processes.	41
3.3	Tail distribution for duration of CA_\equiv with various (x_0, y_0)	47
3.4	Tail distribution for duration of CA_\neq with $r = 1.2$ and various (x_0, y_0)	48
3.5	Tail distribution for duration of CA with various r	50
3.6	Tail distribution for intensity of CA_\equiv with various (x_0, y_0) . The dots are simulation results. The solid lines are the asymptotes from Eq. (3.6).	51
3.7	Tail distribution for intensity of CA with various r	52
3.8	“Delusion of the weakest”: sample paths for different values of r	54
3.9	Scatter plot of duration vs. intensity for CA_\equiv	55
3.10	Conditional average intensity of competitions conditioned on their duration being at least t	55
4.1	Average delay over time in a complete ternary tree with 1093 nodes.	78
4.2	Average delay against network size for a line network.	79
4.3	Average delay against network size for complete quaternary tree.	80
4.4	Average delay against network size for random 3-regular network.	81
A.1	Proof of Claims 1–3 of Appendix A.1.	85
A.2	Proof of Claims 4–5 of Appendix A.1.	86
A.3	Proof of Claim 6 of Appendix A.1.	87
A.4	Proof of Claim 3 of Appendix A.3.	94
A.5	Proof of Claim 4 of Appendix A.3.	95

CHAPTER 1

INTRODUCTION

Most complex systems, be they technological or social, are naturally modeled as networks, or, from a mathematical point of view, graphs, where nodes correspond to discrete entities or components of a system and edges represent dyadic relations between components. Prominent examples are online social networks such as Facebook and Google+, which make visible social ties between individuals. Another example is furnished by the online encyclopedia, Wikipedia, which is a substantiation of the structure of human knowledge. The emergence of these networks has triggered an enormous amount of interest in the scientific community, and considerable work has been devoted to their study.

One line of study focuses on the structural properties of networks themselves. Extensive empirical study has been conducted for a wide range of social and technological networks, focusing on various network characteristics, such as degree distribution, reciprocity and clustering coefficient. Empirical study has led to many interesting discoveries, such as the ubiquitous power-law degree distributions and the small world property. On the theoretical side, a variety of models, either static or dynamic, have been proposed to explain the empirical observations such as the power-law degree distributions and the small world property.

Another line of study, in contrast, focuses on dynamic processes on networks. In this case, networks often serve as platforms for the ongoing processes, which usually involve information dissemination. The focus here is on understanding the interaction of network structural characteristics and dynamic processes. For instance, network

structures can affect, explicitly or implicitly, the efficiency of information dissemination, which may in turn change the network structures, as exemplified by the constant follow and unfollow actions on the Twitter network.

This thesis studies various aspects of networks along the two lines mentioned above. The goal is to advance our understanding of structures and behaviors of large scale complex networks.

In the first part of the thesis, we focus on the interpretation of observed network characteristics. The availability of vast amounts of large scale network data has made it possible to study network characteristics and system behaviors as never before, so we constantly face the problem of properly interpreting empirical observations. Different network properties are commonly studied and interpreted separately. However, they are generally interdependent. Specifying one usually imposes nontrivial constraints on another. Taking such constraints into account provides an additional, and perhaps more appropriate, way of interpreting empirical observations. As a demonstration of the usefulness of this approach, we study the maximum reciprocity, i.e. the maximum percentage of edges with a reciprocal edge, that is realizable by a network with prescribed degree sequence.

In the second part of this thesis, we study network growth dynamics under cumulative advantage, which refers to the “rich-get-richer” phenomenon. An example of such growth dynamics is the Bianconi-Barabási model [12]. The growth process can be viewed as a competition for links, which, directly or indirectly, represent some kind of resource. The evolution over time of the competition, and in particular the change in leadership, is a very intricate process that depends on the interplay of the cumulative advantage effect, individual competitiveness or fitness, and randomness. We focus on the relative leadership between two nodes and characterize the reduced competition dynamics in terms of its duration and intensity.

In the third part of this thesis, we study information dissemination in social networks. The efficiency of information dissemination as characterized by propagation delay is affected by both network topology and user behavior. We analyze the efficiency for both cooperative and selfish user behaviors in various network topologies, and explore the design of incentive mechanisms when cooperation is critical to efficient information dissemination.

1.1 Thesis Contributions

This thesis makes the following main contributions.

- We formulate the *maximum reciprocity problem* that seeks the maximum reciprocity realizable by a network with given degree constraints, and prove its NP-hardness. We provide an upper bound together with necessary conditions and sufficient conditions for achieving the bound. We find that this bound is surprisingly close to the empirical reciprocity in a wide range of real networks, and that there is a surprisingly strong linear relationship between the two. We partially characterize networks with maximum reciprocity by identifying some suboptimal motifs. We demonstrate that a particular type of small suboptimal motifs called 3-paths are the major cause for suboptimality in real networks.
- We characterize the tail distributions of duration and intensity for pairwise competition under cumulative advantage. When the two competitors are equally competitive or fit, we obtain the exact asymptotic distributions. When they are not equally fit, we obtain asymptotic bounds on the distributions. We demonstrate that duration always has a power-law tail irrespective of competitors' fitness, while intensity has either a power-law tail or an exponential tail depending on whether the competitors are equally fit. We observe the struggle-of-

the-fitness phenomenon, where a slight difference in fitness results an extremely heavy tail of duration distribution.

- We characterize for various network topologies the information propagation delay for both cooperative and selfish user behaviors and the corresponding price of stability. We identify topologies where cooperation plays a critical role in efficient information propagation. We propose an incentive mechanism called “plus-one” to coax users into cooperation in such cases, and demonstrate its effectiveness through simulation.

The rest of this thesis is organized as follows. Chapter 2 presents our investigation on reciprocity in networks with degree constraints. Chapter 3 analyzes competition dynamics under cumulative advantage. Chapter 4 explores the efficiency of information dissemination in social networks with limited budget of attention. We conclude in Chapter 5 and discuss some future directions.

CHAPTER 2

RECIPROCITY IN NETWORKS WITH DEGREE CONSTRAINTS

2.1 Introduction

Many complex networks are naturally directed, which endows them with nontrivial structural properties not shared by undirected networks. One such property that has been widely studied is *reciprocity*, which is classically defined as the fraction of edges that are reciprocated, i.e. paired with an edge of the opposite direction. Nontrivial patterns of reciprocity can reveal possible mechanisms of social, biological or other nature that systematically act as organizing principles shaping the observed network topology [31]. Previous work shows that reciprocity plays an important role in many information networks such as email networks [57], the World Wide Web [3] and Wikipedia [81, 82]. It is also shown that major online social networks that are directed in nature, such as Twitter[44, 51], Google+[53], Flickr [56, 18], LiveJournal [78, 56, 32], and YouTube [56], all exhibit a nontrivial amount of reciprocity.

When we try to interpret observed values of reciprocity, we are faced with the problem of assessing the significance of the observation. For instance, the Swedish Wikipedia has reciprocity of 21%. How significant is this? This question is often answered by comparing measured values with the expected value of some null model. One commonly used null model is a random graph with the same number of nodes and edges [57]. An alternative is a random graph with specified degree sequence, as the specific degree sequence is expected to affect reciprocity [79]. Networks are then classified as *reciprocal* or *anti-reciprocal* according to whether the observed reciprocity

is larger or smaller than the expected value [31]. Significant deviation from the expected values suggests the existence of some underlying organizational mechanism at work. For our example of Swedish Wikipedia, the expected reciprocities in both random null models are almost zero, so the Swedish Wikipedia is classified as a reciprocal network. Informative as this might be, comparison with expected values is not the whole story. Is 21% a significant deviation from 0? Can we say that the tendency to reciprocate is strong in this network? The answer might depend on the eye of the beholder. However, if we know for some reason the maximum possible reciprocity is only 28%, then we may safely conclude that 21% is indeed a significant amount of reciprocity. On the other hand, if the maximum is 90%, we might conclude that 21% is not as significant as suggested by the comparison with random null models. In general, knowledge of the extremal values can give a better idea about where the observation lies in the entire spectrum, which can potentially change our conclusion about the significance level of the observation.

Since real social networks often exhibit reciprocities larger than those associated with the random null models, we concern ourselves only with the maximum achievable reciprocity in this chapter. As in the random null models, we may want to retain certain key structural features of the real network when we maximize reciprocity. The particular feature that we choose to preserve here is the joint in- and out-degree sequence, which is a confounding factor in the study of reciprocity [79]. In real networks, in- and out-degrees often serve as proxies for some kind of capacities of the corresponding node. For example, in a file sharing network where edges represent transfers from file sources to downloaders, the in-degree of a node can reflect the available network bandwidth and the out-degree the amount of resource. In a social network where edges point from followers to followees, the in-degree of a node can reflect its fame and popularity and the out-degree its budget of attention. Quite often these capacity constraints are too important to be ignored in the network under con-

sideration. By preserving the degree sequence, we honor these capacity constraints, thus controlling these confounding factors.

Motivated by the above considerations, we study the problem of maximizing reciprocity subject to prescribed joint in- and out-degree constraints. We make the following contributions in this chapter.

- We formulate the *maximum reciprocity problem*. We provide a simple upper bound on reciprocity and prove that it is NP-complete to decide the achievability of the bound. We also identify sufficient conditions for achieving the bound.
- We demonstrate that the upper bound is surprisingly close to the empirical reciprocity in a wide range of real networks, which suggests that the tendency to form reciprocal edges might be much stronger than the observed reciprocity indicates.
- We identify some suboptimal network motifs that can be eliminated to increase reciprocity, thus providing a partial characterization of networks with maximum reciprocity. Based on a particular type of small suboptimal motif called 3-paths, we provide a greedy algorithm GreedyRewire to maximize reciprocity. We demonstrate that 3-paths are the major cause for suboptimality in real networks.
- We find a surprisingly strong linear relationship between the empirical reciprocity and the upper bound across a wide range of real networks. We also find a similar linear relationship between the number of reciprocated edges and the corresponding upper bound on the logarithmic scale.

The rest of the chapter is organized as follows. Section 2.2 introduces the *maximum reciprocity problem*. Section 2.3 proves the NP-hardness of the problem, and provides a simple upper bound for the maximum reciprocity. Section 2.4 identifies patterns of maximum digraphs and provides a greedy algorithm for eliminating

suboptimal motifs. Section 2.5 conducts some empirical study of real networks and Section 2.6 concludes this chapter.

2.2 Graphic Sequences and Maximum Reciprocity Problem

In this section, we first introduce the notion of a *graphic sequence* for undirected graphs and then a *graphic bi-sequence* for directed graphs or digraphs for short, which will be used in the theoretical analysis of Section 2.3. We then formulate the maximum reciprocity problem. Throughout the rest of the chapter, a graph, directed or not, always means a simple graph, i.e. no self-loops or multiple edges are allowed. We will use the terms *node* and *vertex* interchangeably. For directed graphs, an *edge* always means a *directed edge*.

2.2.1 Graphic Degree and Bi-degree Sequences

For an undirected graph $G = (V, E)$, the degree $d_G(v)$ of a node v is the number of edges incident to v . Associated with every graph G is a sequence $\mathbf{d} = \{d_G(v) : v \in V\}$ of its degrees. However, not every sequence of nonnegative integers can be realized by a graph. When it is realizable, the sequence is called *graphic*. More precisely, a sequence of nonnegative integers $\mathbf{d} = (d_1, d_2, \dots, d_n)$ is called *graphic* if there exists a graph G with nodes v_1, v_2, \dots, v_n such that $d_G(v_i) = d_i$ for $i = 1, 2, \dots, n$. The following classical theorem of Erdős and Gallai [27] characterizes graphic sequences.

Theorem 2.2.1 (Erdős-Gallai; Theorem 6.6 in [11]). *A sequence of nonnegative integers $d_1 \geq d_2 \geq \dots \geq d_n$ is graphic if and only if $\sum_{i=1}^n d_i$ is even and*

$$\sum_{i=1}^k d_i \leq k(k-1) + \sum_{i=k+1}^n \min\{d_i, k\}, \quad \text{for } k = 1, 2, \dots, n.$$

The graphicality of a sequence can be tested in linear time using the Erdős-Gallai theorem [40]. It can also be tested using the constructive Havel-Hakimi algorithm in $O(n^2 \log n)$ time [37, 36].

For a digraph $G = (V, E)$, a node has both an in-degree and an out-degree. The in-degree $d_G^-(v)$ of a node v is the number of directed edges coming into v , and the out-degree $d_G^+(v)$ is the number of directed edges going out of v . Associated with every digraph G is a *bi-sequence* $(\mathbf{d}^+, \mathbf{d}^-)$, where $\mathbf{d}^+ = \{d_G^+(v) : v \in V\}$ is the out-degree sequence and $\mathbf{d}^- = \{d_G^-(v) : v \in V\}$ is the in-degree sequence. As in the undirected case, not every bi-sequence of nonnegative integers can be realized by a digraph. A bi-sequence of nonnegative integers $(\mathbf{d}^+, \mathbf{d}^-) = \{(d_1^+, d_2^+, \dots, d_n^+), (d_1^-, d_2^-, \dots, d_n^-)\}$ is called *graphic* if there exists a digraph G with nodes v_1, v_2, \dots, v_n such that $d_G^+(v_i) = d_i^+$ and $d_G^-(v_i) = d_i^-$ for $i = 1, 2, \dots, n$. The Fulkerson-Chen-Anstee theorem is the analog of the Erdős-Gallai theorem for graphic bi-sequences [30, 19, 5].

Theorem 2.2.2 (Fulkerson-Chen-Anstee). *A bi-sequence $\{(d_1^+, \dots, d_n^+), (d_1^-, \dots, d_n^-)\}$ with $d_1^+ \geq d_2^+ \geq \dots \geq d_n^+$ is graphic if and only if $\sum_{i=1}^n d_i^+ = \sum_{i=1}^n d_i^-$ and*

$$\sum_{i=1}^k d_i^+ \leq \sum_{i=1}^k \min\{d_i^-, k-1\} + \sum_{i=k+1}^n \min\{d_i^-, k\}, \quad \text{for } k = 1, 2, \dots, n.$$

The condition of the Fulkerson-Chen-Anstee theorem can be tested in $O(n^2)$ time. The graphicality of bi-sequence can also be tested using the constructive Kleitman-Wang algorithm in $O(n^2 \log n)$ time [48].

2.2.2 Maximum Reciprocity Problem

In this subsection, we formulate the maximum reciprocity problem. For notational simplicity, we henceforth make no distinction between a graph (digraph) and its edge set when no confusion arises.

Given a digraph G , let G_s be the *symmetric* subgraph of G , i.e. $(i, j) \in G_s$ if and only if both $(i, j) \in G$ and $(j, i) \in G$. The reciprocated edges of a digraph G are precisely those of G_s . Thus the number $\rho(G)$ of reciprocated edges in G is given by $\rho(G) = |G_s|$, and the *reciprocity* of G is $r(G) := \rho(G)/|G|$. Note that we use $|G|$ to

denote the number of edges in G and that each pair of reciprocal edges contributes two to $\rho(G)$.

Given a graphic bi-sequence $(\mathbf{d}^+, \mathbf{d}^-)$, let $\mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$ denote the nonempty set of graphs that have $(\mathbf{d}^+, \mathbf{d}^-)$ as their degree bi-sequence. Since the total number of edges is fixed for a given graphic bi-sequence, maximizing $r(G)$ is the same as maximizing $\rho(G)$. The *maximum reciprocity problem* is then to find a digraph G in $\mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$ with maximum $\rho(G)$, i.e.

$$\begin{aligned} & \text{maximize} && \rho(G) \\ & \text{subject to} && G \in \mathcal{G}(\mathbf{d}^+, \mathbf{d}^-). \end{aligned}$$

We denote the maximum value by $\rho(\mathbf{d}^+, \mathbf{d}^-)$ and refer to a digraph G with $\rho(G) = \rho(\mathbf{d}^+, \mathbf{d}^-)$ as a *maximum reciprocity digraph* or *maximum digraph* for short.

2.2.3 Some Notations

We collect here some notations for later reference. Let G denote a generic digraph.

- Let G_a be the anti-symmetric subgraph of G , i.e. $(i, j) \in G_a$ if and only if $(i, j) \in G$ but $(j, i) \notin G$. Note that $G = G_s + G_a$ and $G_s \cap G_a = \emptyset$, i.e. G is the edge disjoint union of G_s and G_a .
- Let G_u be the undirected graph obtained by symmetrizing G , i.e. $(i, j) \in G_u$ if either $(i, j) \in G$ or $(j, i) \in G$.

Let $(\mathbf{d}^+, \mathbf{d}^-)$ be a graphic bi-sequence.

- The min sequence is

$$\mathbf{d}^+ \wedge \mathbf{d}^- = (d_1^+ \wedge d_1^-, d_2^+ \wedge d_2^-, \dots, d_n^+ \wedge d_n^-),$$

where $a \wedge b = \min\{a, b\}$.

- The max sequence is

$$\mathbf{d}^+ \vee \mathbf{d}^- = (d_1^+ \vee d_1^-, d_2^+ \vee d_2^-, \dots, d_n^+ \vee d_n^-),$$

where $a \vee b = \max\{a, b\}$.

- The total number of edges is

$$\varepsilon(\mathbf{d}^+, \mathbf{d}^-) = \sum_i d_i^+ = \sum_i d_i^-.$$

- The total balanced degree is

$$\beta(\mathbf{d}^+, \mathbf{d}^-) = \sum_i d_i^+ \wedge d_i^-,$$

which is the ℓ_1 -norm of the min sequence.

- The total unbalanced degree is

$$\nu(\mathbf{d}^+, \mathbf{d}^-) = \frac{1}{2} \sum_i |d_i^+ - d_i^-|,$$

which is the total variation distance between \mathbf{d}^+ and \mathbf{d}^- . Note that $\varepsilon(\mathbf{d}^+, \mathbf{d}^-) = \beta(\mathbf{d}^+, \mathbf{d}^-) + \nu(\mathbf{d}^+, \mathbf{d}^-)$.

2.3 Hardness Analysis and Bounds

In this section, we first provide an upper bound for the maximum number of reciprocated edges allowed by a graphic bi-sequence. We then prove that the maximum reciprocity problem is NP-hard by showing that it is NP-complete to decide the achievability of the upper bound. Some sufficient conditions for achieving the upper bound are then provided.

2.3.1 Upper Bound for Reciprocity

In this subsection, we first establish a simple upper bound on the maximum number of reciprocal edges in terms of the total balanced degree of the graphic bi-sequence, along with necessary conditions for achieving this upper bound. Some examples are provided to illustrate how the necessary conditions may fail and that they are not sufficient, which provides insight into why the bound is not always tight.

Proposition 2.3.1. *The number of reciprocated edges in any digraph with a given degree bi-sequence cannot exceed the total balanced degree, i.e.*

$$\rho(\mathbf{d}^+, \mathbf{d}^-) \leq \beta(\mathbf{d}^+, \mathbf{d}^-).$$

A necessary condition for equality is that both $\mathbf{d}^+ \wedge \mathbf{d}^-$ and $\mathbf{d}^+ \vee \mathbf{d}^-$ be graphic.

Proof. Let $G \in \mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$ be a maximum digraph. Note that the number of reciprocated edges going out of a node v is at most $d_G^+(v) \wedge d_G^-(v)$. The desired bound is obtained by summing over v .

If equality holds, then G_s , as an undirected graph, has degree sequence $\mathbf{d}^+ \wedge \mathbf{d}^-$, and G_u has degree sequence $\mathbf{d}^+ \vee \mathbf{d}^-$. Thus both $\mathbf{d}^+ \wedge \mathbf{d}^-$ and $\mathbf{d}^+ \vee \mathbf{d}^-$ are graphic. \square

Note that it is possible that neither $\mathbf{d}^+ \wedge \mathbf{d}^-$ nor $\mathbf{d}^+ \vee \mathbf{d}^-$ is graphic. In fact, one sequence can fail to be graphic independently of whether the other is graphic or not, as illustrated by the following examples, where graphic bi-sequences are shown along with the corresponding maximum digraphs.

Example 2.3.2. *In Figure 2.1, neither the min sequence $\mathbf{d}^+ \wedge \mathbf{d}^-$ nor the max sequence $\mathbf{d}^+ \vee \mathbf{d}^-$ is graphic, since they both have odd sums. Here $\rho(\mathbf{d}^+, \mathbf{d}^-) = 2 < \beta(\mathbf{d}^+, \mathbf{d}^-) = 3$.*

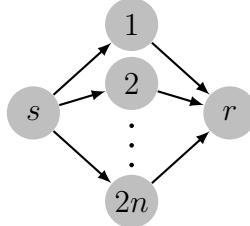
Example 2.3.3. *In Figure 2.2, the min sequence $\mathbf{d}^+ \wedge \mathbf{d}^-$ is graphic, while the max sequence $\mathbf{d}^+ \vee \mathbf{d}^-$ is not. No reciprocity is allowed by this bi-sequence, i.e. $\rho(\mathbf{d}^+, \mathbf{d}^-) =$*



i	(d_i^+, d_i^-)	$d_i^+ \wedge d_i^-$	$d_i^+ \vee d_i^-$
1	(1, 0)	0	1
2	(1, 1)	1	1
3	(0, 2)	0	2
4	(2, 1)	1	2
5	(1, 1)	1	1

Figure 2.1: Graphic bi-sequence with non-graphic max and min sequences.

0, while the upper bound gives $\beta(\mathbf{d}^+, \mathbf{d}^-) = 2n$, so the gap can be arbitrarily large. The only unbalanced nodes s and r have very large unbalanced degrees that cannot be absorbed by themselves, as a consequence of which some, in fact all, balanced degrees have to be used for absorbing unbalanced degrees rather than forming reciprocal edges.



i	(d_i^+, d_i^-)	$d_i^+ \wedge d_i^-$	$d_i^+ \vee d_i^-$
s	($2n$, 0)	0	$2n$
$1 \sim 2n$	(1, 1)	1	1
r	(0, $2n$)	0	$2n$

Figure 2.2: Graphic bi-sequence with graphic min sequence but non-graphic max sequence.

Example 2.3.4. In Figure 2.3, the max sequence $\mathbf{d}^+ \vee \mathbf{d}^-$ is graphic, while the min sequence $\mathbf{d}^+ \wedge \mathbf{d}^-$ is not. As in Example 2.3.3, no reciprocity is allowed here, i.e. $\rho(\mathbf{d}^+, \mathbf{d}^-) = 0$, while the upper bound is $\beta(\mathbf{d}^+, \mathbf{d}^-) = 2n$. The situation is, however, the opposite. Node 0 has too large a balanced degree relative to the number of nodes with nonzero balanced degrees, which is one here. Thus some of the balanced degrees have to be absorbed by the unbalanced degrees.

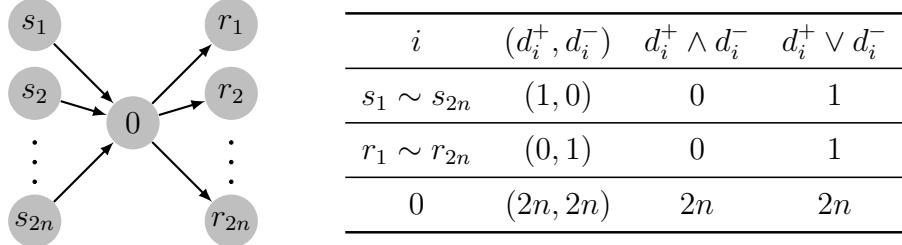


Figure 2.3: Graphic bi-sequence with graphic max sequence but non-graphic min sequence.

The common pattern in Examples 2.3.3 and 2.3.4 is that there are a small number of nodes with extremely large degrees. In the social network context, these nodes correspond to celebrities (node r in Figure 2.3.3), information aggregators (node s in Figure 2.3.3), or middlemen (node 0 in Figure 2.3.4). These large degree nodes often incur inevitable reduction of reciprocity from the upper bound.

The next example shows that the necessary condition in Proposition 2.3.1 is not sufficient.

Example 2.3.5. For the bi-sequence $(d_i^+, d_i^-) = (n - i, i)$, $i = 0, 1, \dots, n$, the upper bound is $\beta(\mathbf{d}^+, \mathbf{d}^-) = \lfloor n/2 \rfloor \cdot \lceil n/2 \rceil$. When n is a multiple of 4, both the max sequence $\mathbf{d}^+ \vee \mathbf{d}^-$ and the min sequence $\mathbf{d}^+ \wedge \mathbf{d}^-$ are graphic. However, $\rho(\mathbf{d}^+, \mathbf{d}^-) = 0$, as the only digraph in $\mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$, of which (i, j) is an edge if and only if $i < j$, has zero reciprocity; see Figure 2.4.

2.3.2 Proof of NP-hardness

We saw in the previous subsection that the upper bound may not be achievable. Unfortunately, the next theorem shows that it is NP-complete to decide whether the upper bound is achievable, which means the maximum reciprocity problem is NP-hard.

Theorem 2.3.6. The decision problem whether $\rho(\mathbf{d}^+, \mathbf{d}^-) = \beta(\mathbf{d}^+, \mathbf{d}^-)$ is NP-complete.

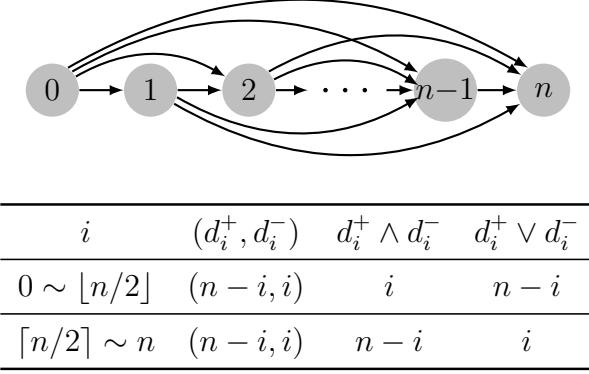


Figure 2.4: Insufficiency of necessary condition in Proposition 2.3.1.

Proof. Note that the problem is the same as the existence problem of a digraph $G \in \mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$ with $\rho(G) = \beta(\mathbf{d}^+, \mathbf{d}^-)$. This problem is in NP, since given a digraph G , we can verify whether $\rho(G) = \beta(\mathbf{d}^+, \mathbf{d}^-)$ in polynomial time. To show that the problem is NP-hard, we adapt the proof of Lemma 5 in [24] by reduction from the 3-color tomography problem, which is shown to be NP-hard therein.

Recall the 3-color tomography problem is as follows. Given nonnegative integral vectors $r^w, r^b \in \mathbb{N}^n$, and $s^w, s^b \in \mathbb{N}^m$ that satisfy

$$r_i^w + r_i^b \leq m, \quad s_j^w + s_j^b \leq n, \quad \text{for } 1 \leq i \leq n, 1 \leq j \leq m,$$

and

$$\sum_i r_i^c = \sum_j s_j^c, \quad \text{for } c \in \{w, b\},$$

decide if (r^w, r^b, s^w, s^b) is feasible, i.e. there exists a matrix M with entries in $\{w, b, g\}$ such that

$$r_i^c = |\{j : M_{ij} = c\}|, \quad s_j^c = |\{M_{ij} = c\}|, \quad \text{for } c \in \{w, b\}.$$

Let (r^w, r^b, s^w, s^b) be an $n \times m$ instance of the 3-color tomography problem. For $1 \leq i \leq n$ and $1 \leq j \leq m$, let

$$\begin{aligned} d_i^+ &= r_i^w + r_i^b + n - 1, & d_{n+j}^+ &= s_j^w, \\ d_i^- &= r_i^w + n - 1, & d_{n+j}^- &= s_j^w + s_j^b. \end{aligned} \tag{2.1}$$

Now we show that the instance (r^w, r^b, s^w, s^b) is feasible if and only if $(\mathbf{d}^+, \mathbf{d}^-)$ is graphic and $\rho(\mathbf{d}^+, \mathbf{d}^-) = \beta(\mathbf{d}^+, \mathbf{d}^-)$, where $\beta(\mathbf{d}^+, \mathbf{d}^-) = n(n-1) + 2 \sum_{i=1}^n r_i^w$.

First assume that M is a solution to the 3-color tomography instance. We construct a digraph G as follows. For $1 \leq i \leq n$ and $1 \leq j \leq m$, let $W_{ij} = 1$ if $M_{ij} = w$, and $B_{ij} = 1$ if $M_{ij} = b$. Let J be an $n \times n$ matrix with all off-diagonal entries equal to 1 and diagonal entries equal to 0. Let the adjacency matrix of G be

$$\begin{pmatrix} J & W + B \\ W^T & 0 \end{pmatrix}.$$

It is straightforward to verify that $G \in \mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$ and $\rho(G) = \beta(\mathbf{d}^+, \mathbf{d}^-)$.

For the reverse direction, assume that $(\mathbf{d}^+, \mathbf{d}^-)$ is graphic and $\rho(\mathbf{d}^+, \mathbf{d}^-) = \beta(\mathbf{d}^+, \mathbf{d}^-)$. Then there exists a digraph $G \in \rho(\mathbf{d}^+, \mathbf{d}^-)$ with $\rho(G) = \beta(\mathbf{d}^+, \mathbf{d}^-)$. Divide the adjacency matrix of G into the following block form

$$G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}.$$

where G_{11} is $n \times n$ and G_{22} is $m \times m$.

Let $\Phi = \sum_{j=1}^n d_j^- - \sum_{i=1}^m d_{n+i}^+$, which, by (2.1), evaluates to $n(n-1)$. On the other hand, $d_j^- = \sum_{k=1}^{n+m} G(k, j)$ and $d_{n+i}^+ = \sum_{k=1}^{n+m} G(n+i, k)$, so

$$\Phi = \sum_{i=1}^n \sum_{j=1}^n G_{11}(i, j) - \sum_{i=1}^m \sum_{j=1}^m G_{22}(i, j) \leq n(n-1) = \Phi,$$

where the inequality follows from the facts that $G_{11}(i, j) \leq 1$, $G_{11}(i, i) = 0$ and $G_{22}(i, j) \geq 0$. Since the equality holds, we must have $G_{11} = J$ and $G_{22} = 0$. Thus

$$\begin{aligned}\rho(G) &= n(n-1) + 2 \sum_{i=1}^n \sum_{j=1}^m G_{12}(i, j)G_{21}(j, i) \\ &\leq n(n-1) + 2 \sum_{i=1}^n \sum_{j=1}^m G_{21}(j, i) \\ &= n(n-1) + 2 \sum_{j=1}^m d_{n+j}^+ = \beta(\mathbf{d}^+, \mathbf{d}^-) = \rho(G).\end{aligned}$$

Since the equality holds, $G_{12}(i, j) \geq G_{21}(j, i)$. Thus $G_{12} = W + B$ and $G_{21} = W^T$ for some $(0, 1)$ -matrices W and B . Let $M_{ij} = w$ if $W(i, j) = 1$, and $M_{ij} = b$ if $B_{ij} = 1$. Then M is a solution to the 3-color tomography instance.

Thus we have shown that it is NP-complete to decide whether $(\mathbf{d}^+, \mathbf{d}^-)$ is graphic and $\rho(\mathbf{d}^+, \mathbf{d}^-) = \beta(\mathbf{d}^+, \mathbf{d}^-)$. Since the graphicality of $(\mathbf{d}^+, \mathbf{d}^-)$ can be tested in quadratic time using the Fulkerson-Chen-Anstee theorem, it must be NP-complete to decide whether $\rho(\mathbf{d}^+, \mathbf{d}^-) = \beta(\mathbf{d}^+, \mathbf{d}^-)$. \square

2.3.3 Sufficient Conditions for Achieving Upper Bound

Given the hardness of the maximum reciprocity problem, we provide some sufficient conditions for achieving the upper bound in Proposition 2.3.1. We start with the following slightly more general theorem, which may be used to lower bound $\rho(\mathbf{d}^+, \mathbf{d}^-)$.

Theorem 2.3.7. $\rho(\mathbf{d}^+, \mathbf{d}^-) \geq 2m$ if there exists a sequence \mathbf{d}^0 such that

- (1). \mathbf{d}^0 is graphic with $\sum_i d_i^0 \geq m$,
- (2). the residual bi-sequence $(\mathbf{d}^+ - \mathbf{d}^0, \mathbf{d}^- - \mathbf{d}^0)$ is also graphic,
- (3). $\Delta < \sqrt{\delta n + (\delta - \frac{1}{2})^2} + \frac{3}{2} - \delta$, where $n = |V_0|$, $\Delta = \bigvee_{i \in V_0} (d_i^+ + d_i^- - d_i^0)$ and $\delta = \bigwedge_{i \in V_0} (d_i^+ + d_i^- - d_i^0)$, with $V_0 = \{i : d_i^+ \vee d_i^- > 0\}$.

This theorem is analogous to Theorem 2.2 in [16], which deals with packing two graphic sequences for undirected graphs. Theorem 2.3.7 deals with packing a graphic sequence \mathbf{d}^0 for undirected graphs and a graphic bi-sequence $(\mathbf{d}^+ - \mathbf{d}^0, \mathbf{d}^- - \mathbf{d}^0)$ for digraphs. The proof is deferred to Appendix A.1.

Applying Theorem 2.3.7 with $\mathbf{d}^0 = \mathbf{d}^+ \wedge \mathbf{d}^-$, we obtain the following sufficient conditions for achieving the upper bound in Proposition 2.3.1.

Corollary 2.3.8. $\rho(\mathbf{d}^+, \mathbf{d}^-) = \beta(\mathbf{d}^+, \mathbf{d}^-)$ if the following conditions hold,

- (1). $\mathbf{d}^+ \wedge \mathbf{d}^-$ and $(\mathbf{d}^+ - \mathbf{d}^+ \wedge \mathbf{d}^-, \mathbf{d}^- - \mathbf{d}^+ \wedge \mathbf{d}^-)$ are graphic;
- (2). $\Delta < \sqrt{\delta n + (\delta - \frac{1}{2})^2} + \frac{3}{2} - \delta$, where $n = |V_0|$, $\Delta = \bigvee_{i \in V_0} (d_i^+ \vee d_i^-)$ and $\delta = \bigwedge_{i \in V_0} (d_i^+ \vee d_i^-)$, with $V_0 = \{i : d_i^+ \vee d_i^- > 0\}$.

Note that Δ is the maximum of either the in- or out-degrees. Putting an upper bound on Δ rules out extremely large degrees, which are the trouble makers in the examples of Section 2.3.1. However, in most real networks, we have $\delta = 1$, so the sufficient condition essentially requires $\Delta < \sqrt{n}$, which, unfortunately, usually fails to hold. In fact, it fails for most networks studied in Section 2.5.

2.4 Patterns in Maximum Digraphs

In this section, we identify some structural patterns of maximum digraphs, or equivalently, the associated suboptimal structures that contribute to the loss in reciprocity that is *not* imposed by the degree bi-sequence. We first look at some small suboptimal motifs and provide a greedy algorithm to eliminate them. We then show some more complicated structural patterns of maximum digraphs and demonstrate how they can help us pin down the maximum digraphs in some special cases.

Throughout this section, a cycle or a path always refers to a directed cycle or directed path, i.e. the edges must be all in the same direction as we follow the cycle or path. We also require that the edges be distinct. On the other hand, the vertices

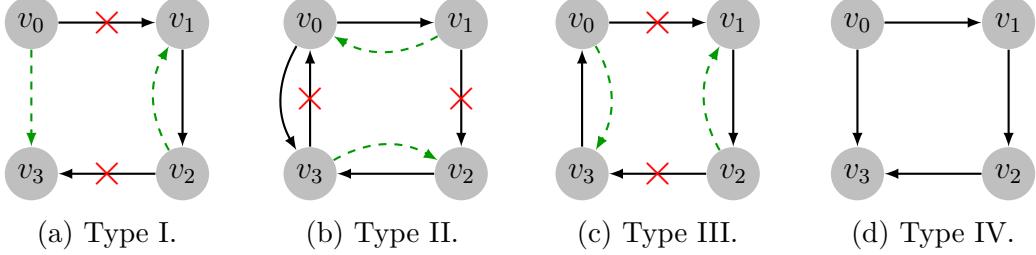


Figure 2.5: Different types of 3-paths with corresponding rewirings. The edges marked by red crosses are to be rewired into the dashed green edges.

are not necessarily distinct. When the vertices are distinct, we say the path or cycle is elementary.

2.4.1 Small Suboptimal Motifs

In this subsection, we focus on a particular type of small motifs that we call 3-paths, the nonexistence of which also guarantees the nonexistence of many larger scale suboptimal structures. As we will see in Section 2.5, elimination of such suboptimal motifs brings reciprocity close to the corresponding upper bound for a variety of real world networks.

Given a digraph G , we call an elementary path of length 3, $\pi = (v_0, v_1, v_2, v_3)$, a 3-path if $(v_i, v_{i+1}) \in G_a$ for $i = 0, 1, 2$, i.e., π consists entirely of unreciprocated edges. We further classify 3-paths into the following four types according to the connectivity between v_0 and v_3 ,

- (I). $(v_0, v_3) \notin G_u$, i.e. there is no edge between v_0 and v_3 ;
- (II). $(v_0, v_3) \in G_s$;
- (III). $(v_3, v_0) \in G_a$, i.e. $(v_0, v_1, v_2, v_3, v_0)$ is a 4-cycle;
- (IV). $(v_0, v_3) \in G_a$.

As shown in Figure 2.5, 3-paths of Types I, II and III are suboptimal and can be rewired locally to increase reciprocity. We say a digraph is *3-path optimal* if it has no 3-path of Type I, II or III. Note that when viewed as a transformation on G_a , the rewiring procedure in Figure 2.5 simply eliminates 4-cycles (Type III), and replaces open 3-paths by a shortcut from its first vertex to its last vertex if such a shortcut does not yet exist (Types I and II). Thus each rewiring increases the number of reciprocated edges by either 2 or 4, and we have the following

Lemma 2.4.1. *A maximum digraph is 3-path optimal.*

Given a digraph G , we can greedily rewire all 3-paths to get a lower bound on the maximum reciprocity allowed by the degree bi-sequence of G . The resulting greedy algorithm is shown in Algorithm 1.

Algorithm 1 GreedyRewire

Input: $G = (V, E)$

- 1: $S \leftarrow V$
- 2: **while** $S \neq \emptyset$ **do**
- 3: pick $v_0 \in S$
- 4: **if** there exists non-Type IV 3-path $\pi = (v_0, v_1, v_2, v_3)$ **then**
- 5: $G \leftarrow \text{Rewire}(\pi)$
- 6: $S \leftarrow S \cup \{v_1, v_2\}$
- 7: **else**
- 8: $S \leftarrow S - \{v_0\}$
- 9: **end if**
- 10: **end while**
- 11: **return** G

Proposition 2.4.2 guarantees that Algorithm 1 eliminates all 3-paths of Types I, II and III. The proof is deferred to Appendix A.2.

Proposition 2.4.2. *Algorithm 1 returns a 3-path optimal digraph.*

Note that depending on how v_0 and π are picked, Algorithm 1 can return different 3-path optimal graphs. Although there is no theoretical guarantee, we will see in Section 2.5 that reciprocities of 3-path optimal digraphs returned by Algorithm 1 are

very close to the corresponding upper bounds and hence close to the maxima as well. The next subsection shows that 3-path optimality precludes many other suboptimal structures, which partially explains why Algorithm 1 works pretty well in practice.

2.4.2 Properties of Maximum Digraphs

In this subsection, we consider additional suboptimal structures that are more complicated than 3-paths. Some of these structures are automatically eliminated by Algorithm 1, while others require extra attention. We will state the results as properties of maximum digraphs. Any violation of the stated properties yields a suboptimal structure.

2.4.2.1 3-path Optimal Digraphs

We first consider some properties of 3-path optimal digraphs, which, by Lemma 2.4.1, are also properties of maximum digraphs. All these properties involve only unreciprocated edges. Note that any suboptimal structures that violate these properties are automatically eliminated by Algorithm 1. Let G denote a 3-path optimal digraph throughout this subsection. Recall that in a 3-path optimal digraph, the only possible 3-path is of Type IV.

Lemma 2.4.3 shows that the unreciprocated edges of a 3-path optimal digraph cannot form any elementary path of odd length without a shortcut. As a result, for any two vertices u and v , either there is no path from u to v in G_a , or there is such a path of length at most 2.

Lemma 2.4.3. *If $\pi = (v_0, v_1, \dots, v_{2p+1})$ is an elementary path of odd length in G_a , then $(v_0, v_{2p+1}) \in G_a$.*

Proof. We use induction on p . If $p = 0$, then $(v_0, v_1) \in G_a$ by assumption. If $p = 1$, then π is a 3-path of Type IV and hence $(v_0, v_3) \in G_a$. Now consider $p \geq 2$. We have

$(v_0, v_{2p-1}) \in G_a$ by the induction hypothesis. Then $(v_0, v_{2p-1}, v_{2p}, v_{2p+1})$ is a 3-path of Type IV. Thus $(v_0, v_{2p+1}) \in G_a$. \square

Lemma 2.4.4 shows that the anti-symmetric subgraph of a 3-path optimal digraph is almost cycle free. We can obtain a directed acyclic graph from it by removing an edge from each 3-cycle.

Lemma 2.4.4. *The only possible cycles in G_a are 3-cycles, and any two distinct 3-cycles must be vertex disjoint.*

Proof. We first prove that two distinct 3-cycles must be vertex disjoint by contradiction. Suppose they share at least one vertex v_0 . Let the cycles be $C_0 = (v_0, v_1, v_2, v_0)$ and $C_1 = (v_0, v_3, v_4, v_0)$. Note that $v_1 \neq v_4$ and $v_2 \neq v_3$, as all edges are in G_a . Since C_0 and C_1 are distinct, we must have either $v_1 \neq v_3$ or $v_2 \neq v_4$. Without loss of generality, assume $v_1 \neq v_3$. Then (v_1, v_2, v_0, v_3) is a 3-path of Type IV, so $(v_1, v_3) \in G_a$. But then (v_1, v_3, v_4, v_0) is a 3-path of Type III, which is impossible. Therefore, C_0 and C_1 must be vertex disjoint.

Next we prove there are no elementary k -cycles for $k \geq 4$. Suppose there is such a cycle $(v_0, v_1, \dots, v_{k-1}, v_0)$. If k is even, $(v_1, v_{k-2}) \in G_a$ by Lemma 2.4.3. But $(v_0, v_1, v_{k-2}, v_{k-1})$ is a 3-path of Type III, which is impossible. If k is odd, then $(v_0, v_{k-2}), (v_1, v_{k-1}) \in G_a$ again by Lemma 2.4.3. But then (v_0, v_1, v_{k-1}, v_0) and $(v_0, v_{k-2}, v_{k-1}, v_0)$ are two distinct 3-cycles with two common vertices, which is again impossible.

Finally, suppose there is a non-elementary cycle. We can decompose it into several distinct elementary cycles, all of which must be 3-cycles by the previous paragraph. But then we have distinct 3-cycles that are not vertex disjoint, which is impossible. Therefore, there are no k -cycles for $k \geq 4$. \square

Although 3-path optimality does not preclude 3-cycles, they are unlikely to exist in 3-path optimal graphs obtained from real world networks using Algorithm 1, as

Lemma 2.4.5 requires that the vertices of a 3-cycle in such graphs have exactly the same connectivity to every vertex outside the 3-cycle, which is extremely unlikely, especially in large graphs.

Lemma 2.4.5. *For a 3-cycle C in G_a and any vertex v not in C , if there is a path π in G_a that connects v and C , then there is an edge of G_a between v and each vertex of C , all in the same direction as π (i.e. from v to C or from C to v).*

Proof. Let $C = (v_0, v_1, v_2, v_0)$. Without loss of generality, assume π is from v to v_0 and has odd length. Successive application of Lemma 2.4.3 to the paths π , (v, v_0, v_1, v_2) and then (v, v_2, v_0, v_1) , we obtain $(v, v_0) \in G_a$, $(v, v_2) \in G_a$ and $(v, v_1) \in G_a$ in the same order. \square

2.4.2.2 Maximum Digraphs

In this subsection, we consider some properties of maximum digraphs that are not direct consequences of 3-path optimality. The associated suboptimal structures may be left intact by Algorithm 1 and require extra attention. Throughout this subsection, let G^* denote a maximum digraph with a given bi-sequence $(\mathbf{d}^+, \mathbf{d}^-)$, i.e. $G^* \in \mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$ and $\rho(G^*) = \rho(\mathbf{d}^+, \mathbf{d}^-)$.

We know from Lemma 2.4.4 that large cycles involving only unreciprocated edges are suboptimal structures, but certain cycles of even length that contains reciprocated edges are also suboptimal. In particular, we have the following

Lemma 2.4.6. *Let C be an even cycle in $H \in \mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$. If any two edges in $C \cap H_s$ are separated by an odd number of edges in C , then there exists $H' \in \mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$ with $\rho(H') = \rho(H) + |C_a| - 2|C_a \cap H_s|$, where C_a is the anti-symmetric part of C , i.e. $C_a = \{(i, j) \in C : (j, i) \notin C\}$.*

Note that $C \cap H_a \subset C_a$ but it is not necessarily true that $C_a = C \cap H_a$. The two edges $(3, 4)$ and $(5, 0)$ in Figure 2.6(a) are in C_a but not in $C \cap H_a$. Any cycle satisfying

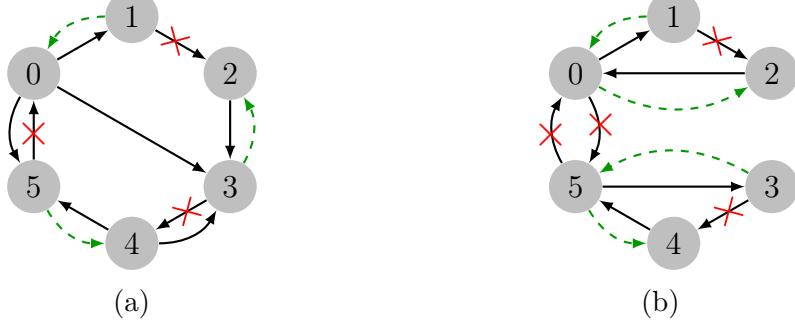


Figure 2.6: Suboptimal even cycle with reciprocated edges. Reciprocity can be increased by rewiring the edges marked by red crosses into the dashed green edges.

the conditions in Lemma 2.4.6 is suboptimal if it has more anti-symmetric edges than symmetric ones. The cycles $(0, 1, 2, 3, 4, 5, 0)$ in Figure 2.6(a) and $(0, 1, 2, 0, 5, 3, 4, 5, 0)$ in Figure 2.6(b) are two such examples. Note that these two cycles are not automatically eliminated by Algorithm 1.

Proof of Lemma 2.4.6. Let $C = (v_0, v_1, \dots, v_{2p-1}, v_{2p} = v_0)$, where the vertices are labeled such that $(v_{2p-1}, v_0) \in H_s$ if $C \cap H_s \neq \emptyset$. Note that the vertices may not be distinct. Note also that we must have $(v_{2k}, v_{2k+1}) \in H_a$ for all k . If $(v_{2k}, v_{2k+1}) \in H_s$ for some k , then the number of edges in C between (v_1, v_2) and (v_{2k}, v_{2k+1}) would be $2k - 2$, contradicting the assumption that any two edges in $C \cap H_s$ are separated by an odd number of edges in C . As illustrated in Figure 2.6, let

$$H' = H - \{(v_{2i-1}, v_{2i})\}_{i=1}^p + \{(v_{2i-1}, v_{2i-2})\}_{i=1}^p.$$

Since $(v_{2i-2}, v_{2i-1}) \in H_a$, we have $(v_{2i-1}, v_{2i-2}) \notin H$ and hence $H' \in \mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$. Note that the edges in $C \cap H_a$ are either absent from H' or in H'_s , so $H_a - H'_a = C \cap H_a = C_a \cap H_a$. On the other hand, all edges in $C \cap H_s$ are removed from H' , so $H'_a - H_a = C_a \cap H_s$. Thus by going from H to H' , we eliminated $|C_a \cap H_a|$ unreciprocated edges while creating $|C_a \cap H_s|$ new ones. Therefore, $\rho(H') = \rho(H) - |C_a \cap H_s| + |C_a \cap H_a|$.

Since $|C_a| = |C_a \cap H_s| + |C_a \cap H_a|$, the desired conclusion follows by eliminating $|C_a \cap H_a|$. \square

Lemma 2.4.7 specifies how multiple 3-cycles should be connected in maximum digraphs. If we collapse each 3-cycle into a single vertex by contracting its edges, the subgraph of G_a^* induced by these vertices will have the structure in Figure 2.4. Therefore, while the existence of multiple 3-cycles is already very unlikely in 3-path optimal digraphs, it is even less likely in maximum digraphs with degree bi-sequences of real world networks.

Lemma 2.4.7. *The set of all distinct 3-cycles of G_a^* can be linearly ordered as C_0, C_1, \dots, C_m such that there are 9 edges of G_a^* going from C_i to C_j for all $0 \leq i < j \leq m$.*

Proof. Consider two distinct 3-cycles $C = (v_0, v_1, v_2, v_0)$ and $C' = (w_0, w_1, w_2, w_0)$. There cannot exist a pair of edges from G_s^* that connect C and C' ; otherwise, say $(v_0, w_0) \in G_s^*$, the cycle $(v_0, v_1, v_2, v_0, w_0, w_1, w_2, w_0, v_0)$ would be suboptimal by Lemma 2.4.6. On the other hand, there must be at least one edge between C and C' ; otherwise, replacing C_i and C_j by the three pairs of edges $\{(v_i, w_i), (w_i, v_i)\}_{i=0}^2$ would increase the reciprocity. Without loss of generality, assume $(v_0, w_0) \in G_a^*$. It then follows from Lemma 2.4.5 that $(v_i, w_j) \in G_a^*$ for all $i, j \in \{0, 1, 2\}$. By Lemma 2.4.4, such edges cannot be part of any cycle. Therefore, we can sort the 3-cycles topologically and label them in the desired way. \square

The next lemma complements Lemma 2.4.3 by specifying connection patterns of elementary paths of even length.

Lemma 2.4.8. *Let $\pi = (v_0, v_1, \dots, v_{2p})$ be an elementary path of even length $2p \geq 4$ in G_a^* , $E_0 = \{(v_{2i}, v_{2j}) : i \neq j\}$ and $E_1 = \{(v_{2i-1}, v_{2j-1}) : i \neq j\}$. If $(v_0, v_{2p}) \notin G_a^*$, then G^* either has all the edges in E_0 but none in E_1 , or vice versa.*

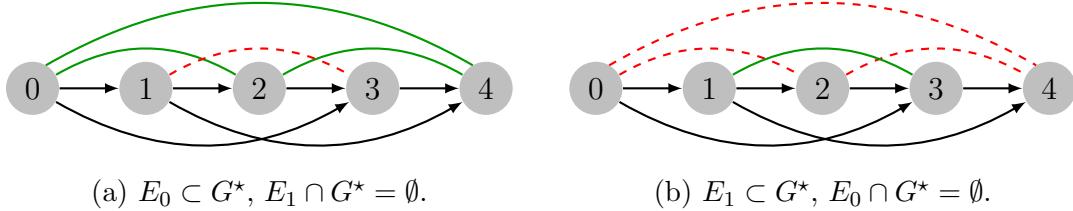


Figure 2.7: Patterns of even paths in maximum digraphs. Each undirected solid edge represents a pair of reciprocated edges in G^* . Each dashed edge represents a pair of edges that are both missing in G^* .

Proof. See Appendix A.3. □

Figure 2.7 shows both possibilities for an elementary path of length 4. The shortcuts required by Lemma 2.4.3 are also shown. The red dashed edges represent those that cannot coexist with the green edges in a maximum digraph. Some suboptimal structures that violate Lemma 2.4.8 cannot be automatically eliminated by Algorithm 1. For example, if the pair of edges between the vertices 0 and 2 are missing from Figure A.4, the resulting suboptimal digraph will be left intact by Algorithm 1.

2.4.3 Some Examples

In this subsection, we illustrate how the structural patterns of the previous subsection may be used to pin down the maximum digraph in some special cases. Here G^* always denotes a maximum digraph.

Proposition 2.4.9 shows that when the bi-sequence is perfectly balanced, the maximum digraph achieves perfect or near-perfect reciprocity. Therefore, any unfulfilled reciprocity must be due to the lack of effort to form reciprocal edges rather than due to the fundamental limit imposed by the bi-sequence itself.

Proposition 2.4.9. *Suppose $(\mathbf{d}^+, \mathbf{d}^-)$ is perfectly balanced, i.e. $\nu(\mathbf{d}^+, \mathbf{d}^-) = 0$.*

(1). *If $\varepsilon(\mathbf{d}^+, \mathbf{d}^-)$ is even, then $\rho(\mathbf{d}^+, \mathbf{d}^-) = \varepsilon(\mathbf{d}^+, \mathbf{d}^-)$.*

(2). *If $\varepsilon(\mathbf{d}^+, \mathbf{d}^-)$ is odd, then $\rho(\mathbf{d}^+, \mathbf{d}^-) = \varepsilon(\mathbf{d}^+, \mathbf{d}^-) - 3$, and G_a^* consists of a 3-cycle.*

Proof. Since $\nu(\mathbf{d}^+, \mathbf{d}^-) = 0$, we have $d_i^+ = d_i^-$ for all i . Thus any edge of G_a^* must be contained in a cycle of length at least 3 in G_a^* . By Lemma 2.4.4, the length of such a cycle is exactly 3. By Lemma 2.4.7, there is at most one such cycle in G_a^* . Thus G_a^* is either empty or a 3-cycle. Since $\rho(G^*)$ must be even, the former case corresponds to even $\varepsilon(\mathbf{d}^+, \mathbf{d}^-)$ and the latter odd $\varepsilon(\mathbf{d}^+, \mathbf{d}^-)$. \square

The next proposition shows that when the bi-sequence is slightly unbalanced, the number of possible values of $\rho(\mathbf{d}^+, \mathbf{d}^-)$ increases. This sheds some light on why the maximum reciprocity problem is so difficult. As the total unbalanced degree increases, the number of possibilities is expected to explode.

Proposition 2.4.10. *Suppose $(\mathbf{d}^+, \mathbf{d}^-)$ is slightly unbalanced with $\nu(\mathbf{d}^+, \mathbf{d}^-) = 1$, $d_0^+ - d_0^- = 1$ and $d_1^- - d_1^+ = 1$.*

(1). *If $\varepsilon(\mathbf{d}^+, \mathbf{d}^-)$ is even, then the gap $\varepsilon(\mathbf{d}^+, \mathbf{d}^-) - \rho(\mathbf{d}^+, \mathbf{d}^-)$ is either 2 or 4. When the gap is 2, the two edges in G_a^* form a 2-path from 0 to 1. When the gap is 4, G_a^* is the vertex disjoint union of $\{(0, 1)\}$ and a 3-cycle.*

(2). *If $\varepsilon(\mathbf{d}^+, \mathbf{d}^-)$ is odd, then the gap $\varepsilon(\mathbf{d}^+, \mathbf{d}^-) - \rho(\mathbf{d}^+, \mathbf{d}^-)$ is either 1 or 5. When the gap is 1, $G_a^* = \{(0, 1)\}$. When the gap is 5, G_a^* is the vertex disjoint union of a 2-path from 0 to 1 and a 3-cycle.*

Proof. Note that there must be a path from 0 to 1 in G_a^* . Let π be the shortest path from 0 to 1 in G_a^* . All edges in $G_a^* - \pi$, if there is any, must be contained in a cycle in G_a^* . By Lemma 2.4.4, G_a^* can only have 3-cycles. If G_a^* had more than one 3-cycles, Lemma 2.4.7 would require that there be at least 9 edges in G_a^* that are not contained in any cycle, all of which must be in π . Lemma 2.4.3 shows that π has either one or two edges. Therefore, $G_a^* - \pi$ is either empty or has one 3-cycle. By Lemma 2.4.5, π and the 3-cycle, if there is one, must be vertex disjoint. Since $|\pi| \in \{1, 2\}$, and $|G_a^* - \pi| \in \{0, 3\}$, it follows that $\varepsilon(\mathbf{d}^+, \mathbf{d}^-) - \rho(G^*) = |G_a^*| = |\pi| + |G_a^* - \pi| \leq 2 + 3 = 5$.

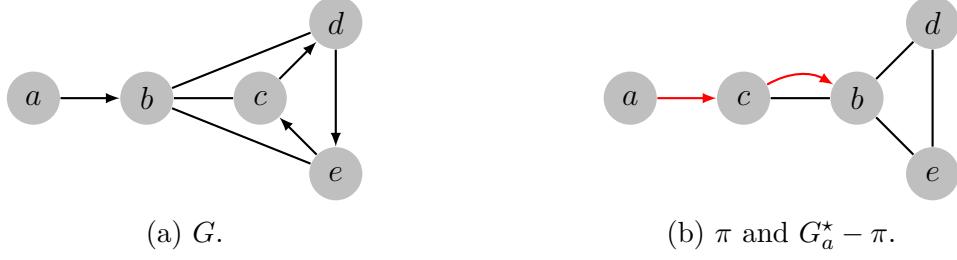


Figure 2.8: Example 2.4.11.

Note that $\rho(G^*)$ is even. If $\varepsilon(\mathbf{d}^+, \mathbf{d}^-)$ is even, then $\varepsilon(\mathbf{d}^+, \mathbf{d}^-) - \rho(G^*)$ is equal to $|\pi| = 2$ or $\alpha(G^*) = |\pi| + |G_a - \pi| = 1 + 3 = 4$. If $\varepsilon(\mathbf{d}^+, \mathbf{d}^-)$ is odd, then $\varepsilon(\mathbf{d}^+, \mathbf{d}^-) - \rho(G^*)$ is equal to $|\pi| = 1$ or $|\pi| + |G_a^* - \pi| = 2 + 3 = 5$. \square

It is easy to come up with examples where the gaps are 1 and 2, respectively. The next examples shows that the other two cases are also possible.

Example 2.4.11. Let $(\mathbf{d}^+, \mathbf{d}^-) = \{(1, 3, 2, 2, 2), (0, 4, 2, 2, 2)\}$. Figure 2.8(a) shows a realization G of this bi-sequence, where each undirected edge represents a pair of edges in opposite directions. Note that $\rho(G) = \varepsilon(\mathbf{d}^+, \mathbf{d}^-) - 4$. We claim that $\rho(G) = \rho(\mathbf{d}^+, \mathbf{d}^-)$. If not, then $\rho(G^*) = \varepsilon(\mathbf{d}^+, \mathbf{d}^-) - 2$ by Proposition 2.4.10, and the two edges in G_a^* form a 2-path π from a to b . Since c, d, e have the same in- and out-degrees and hence are equivalent, we may assume without loss of generality that $\pi = (a, c, b)$. Thus $G_a^* - \pi$ is symmetric and corresponds to a simple graph with degree sequence $\hat{\mathbf{d}} = \{0, 3, 1, 2, 2\}$. There is only one simple graph with this degree sequence, which is shown by the black edges in Figure 2.8(b). When we superimpose π and $G_a^* - \pi$, there are two edges from (c, b) , and hence $G^* \notin \mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$, a contradiction.

Example 2.4.12. Let $(\mathbf{d}^+, \mathbf{d}^-) = \{(1, 0, 4, 2, 2, 2), (0, 1, 4, 2, 2, 2)\}$. Figure 2.9 shows a realization G of this bi-sequence, where each undirected edge represents a pair of edges in opposite directions. Note that $\rho(G) = \varepsilon(\mathbf{d}^+, \mathbf{d}^-) - 5$. Since the sequence $\mathbf{d}^+ \wedge \mathbf{d}^- = \{0, 0, 4, 2, 2, 2\}$ is not graphic, Proposition 2.3.1 shows that $\rho(\mathbf{d}^+, \mathbf{d}^-) <$

$\varepsilon(\mathbf{d}^+, \mathbf{d}^-) = 1$. Thus Proposition 2.4.10 yields $\rho(G) = \rho(\mathbf{d}^+, \mathbf{d}^-)$. In fact, G is the only element of $\mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$.

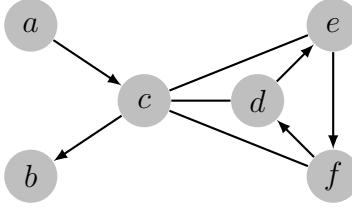


Figure 2.9: Example 2.4.12.

2.5 Empirical Study

In this section, we conduct an empirical analysis of real networks by comparing the observed values of reciprocity against the upper bounds. We also look at the lower bounds on maximum reciprocities given by Algorithm 1.

2.5.1 Datasets

The networks that we analyze include major online social networks (OSN) that are directed in nature [56, 51, 35, 77, 52]. For the purpose of comparison, we have also included other types of networks: biological networks [73, 74, 68, 76, 70, 80], communication networks [52], product co-purchasing networks [52], web graphs [52], Wikipedias [1], software call graphs [75, 66], and P2P networks [52]. All the datasets except for Wikipedias are already converted into graph representations by other researchers and the descriptions for the datasets can be found at the cited sources. For Wikipedias, each node represents a page. Only article pages, i.e. pages with namespace ID 0, are included. Pages that redirect to the same page are represented as a single node corresponding to the destination page. There is an edge from node A to node B if there is at least one hyperlink from page A to page B . Multiple edges and self-loops are discarded. Some basic statistics of the networks can be found in Appendix B.

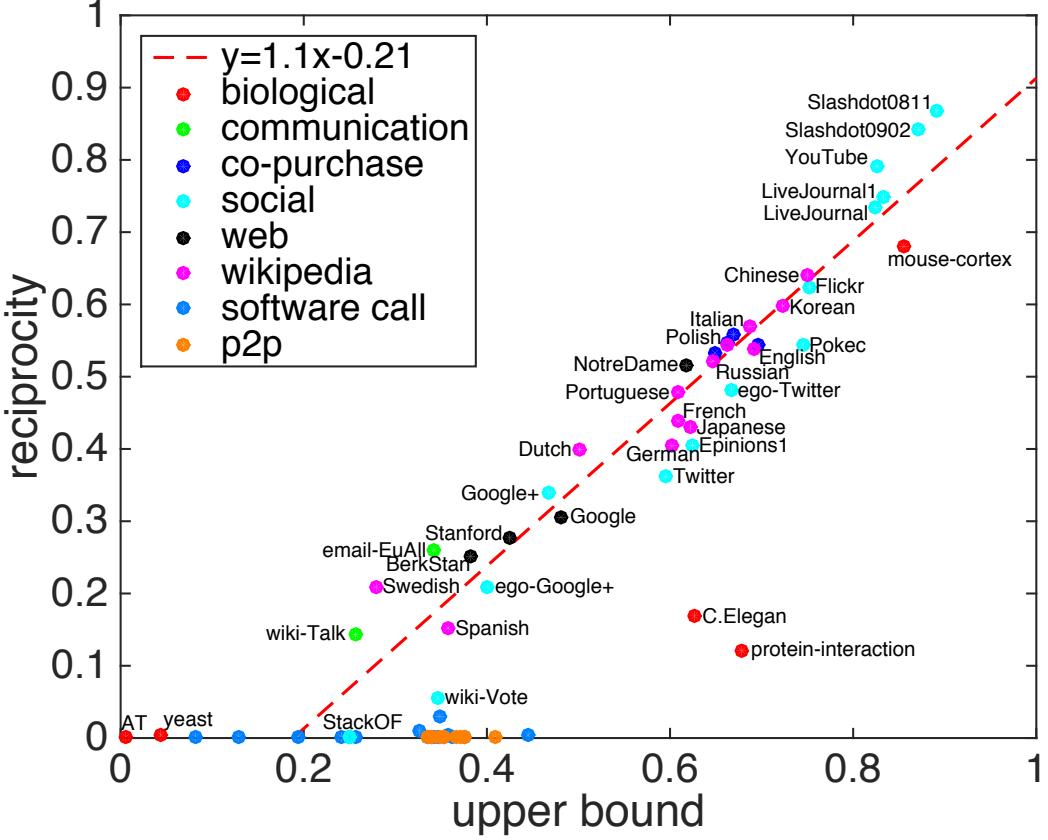


Figure 2.10: Scatter plot of empirical reciprocity versus upper bound. Regression line was fitted without data points for biological, P2P and software call networks.

2.5.2 Empirical Reciprocity vs. Upper Bound

Figure 2.10 shows the scatter plot of empirical reciprocities against the corresponding upper bounds. Here the upper bound is normalized by the number of edges, i.e., it is the ratio $\beta(\mathbf{d}^+, \mathbf{d}^-)/\varepsilon(\mathbf{d}^+, \mathbf{d}^-)$. Note that the reciprocity values vary widely, ranging from 0 for the peer-to-peer network Gnutella to 90% for the online social network Slashdot. There is even a fair amount of variation within the categories of biological, social and Wikipedia networks. In general, social networks and Wikipedia networks tend to have high reciprocity, while software call networks tend to have low reciprocity. Note the strong linear correlation between empirical reciprocity and the upper bound. This is a little bit surprising, especially for the social networks, in view of the large variations in reciprocity. Related to Figure 2.10 is the scatter

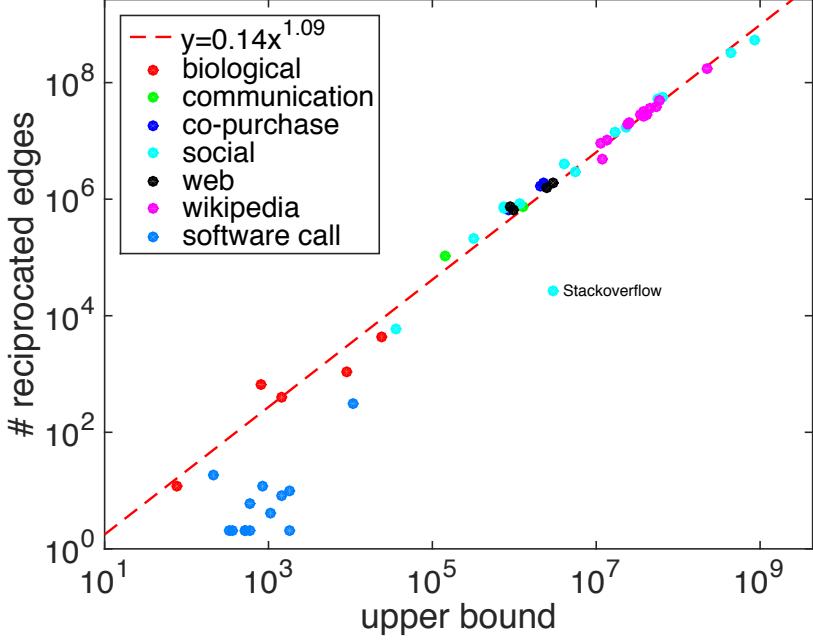


Figure 2.11: Scatter plot of number of reciprocated edges versus upper bound. Regression line was fitted in log scale, without data points for software call networks.

plot in Figure 2.11 of number of reciprocated edges against the unnormalized bound $\beta(\mathbf{d}^+, \mathbf{d}^-)$. There the linear relationship in log-log scale is even more apparent, with biological networks being also around the regression line. These linear relationships suggest that there might exist some universal mechanism that works across different domains.

Despite the wide variation in reciprocity, the ratio between the empirical reciprocity and the normalized upper bound has a much narrower range as shown by the box plots for the ratios in Figure 2.12.

Note that the ratios are close to zero for the P2P network Gnutella and software call graphs. The Gnutella exhibits zero reciprocity, far away from the upper bounds that are above 30%. This is probably because Gnutella implements an indirect reciprocity mechanism. The low reciprocity for software call graphs is not surprising, as software codes are usually designed to work in a hierarchical manner. The case for biological networks are more complicated, as the four biological networks considered

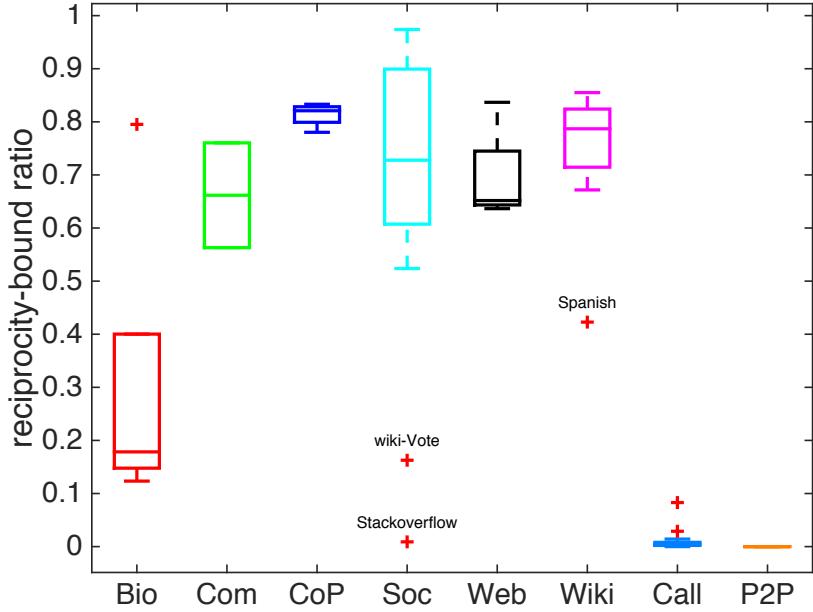


Figure 2.12: Box plot of reciprocity-bound ratio for different network categories.

here are actually of quite different natures. For example, the C. Elegan neural network and the mouse cortex network are both neural networks, but the former is at the neuron level while the latter is at a coarser level of cortical regions. One can speculate that both the low reciprocity in C. Elegan neural network and the high reciprocity in the mouse cortex network are due to biological reasons. However, we do not know if this behavior is a norm or an exception due to the lack of data for similar networks.

In all categories other than biological, software call and P2P networks, the ratios are above 50% with only three exceptions: the wiki-Vote network, the Stack Overflow Q&A network, and the Spanish Wikipedia. Although we have classified the Stack Overflow Q&A network as a social network, it differs from typical social networks. The low reciprocity suggests that there is a hierarchy of expertise. What is more interesting is the wiki-Vote network and the Spanish Wikipedia, as their behaviors deviate from those of other networks of the same category, which suggests that there might be something unusual about them that is worthy of scientific study. Apart from the three outliers, all other networks in these categories actually achieve a significant

fraction of the possible reciprocity suggested by the upper bound. This means that modulo the degree constraints, the tendency to reciprocate is much stronger than the empirical reciprocity alone might have suggested. Prominent examples include the web graphs, the Swedish Wikipedia and the Google+ network, whose reciprocities are not very high in absolute value but quite high relative to the bound. This suggests that when we study these networks, it might be more meaningful to ask the question why there is such large imbalance in degrees than to ask the question why the tendency to reciprocate is low.

2.5.3 Reciprocity of 3-path Optimal Digraphs

In this subsection, we look at 3-path optimal digraphs returned by Algorithm 1. Note that the reciprocity of such a digraph provides a lower bound on the maximum reciprocity of the corresponding degree bi-sequence.

Figure 2.13 shows the scatter plot of the reciprocities of the 3-path optimal digraphs against the corresponding upper bounds. Figure 2.14 shows the box plots of their ratios. Note that the reciprocities of 3-path optimal digraphs are close to the upper bounds, especially for communication, co-purchasing, social and Wikipedia networks. This means that the maximum reciprocities are also close to the upper bounds. Therefore, for the degree bi-sequences of those real networks, the fundamental limit that they impose on reciprocity is largely summarized by the upper bounds, and the major source of loss in reciprocity is the existence of 3-paths of Types I, II and III. Thus in practice Algorithm 1 usually suffices for approximating maximum reciprocities and we do not need to worry much about the more complicated suboptimal structures in Section 2.4.2.2.

Finally, recall from Section 2.4.2.1 that the existence of 3-cycles in a 3-path optimal digraph requires some specific structures. These structures are usually too special to occur in practice, so 3-cycles are unlikely to exist in 3-path optimal digraphs. This

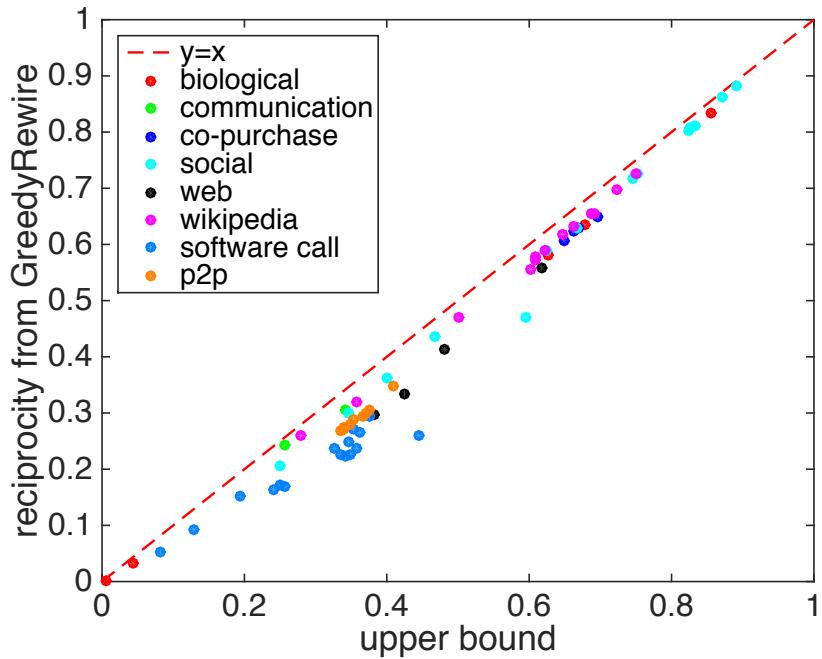


Figure 2.13: Scatter plot of reciprocity of the 3-path optimal digraph returned by Algorithm 1 versus upper bound.

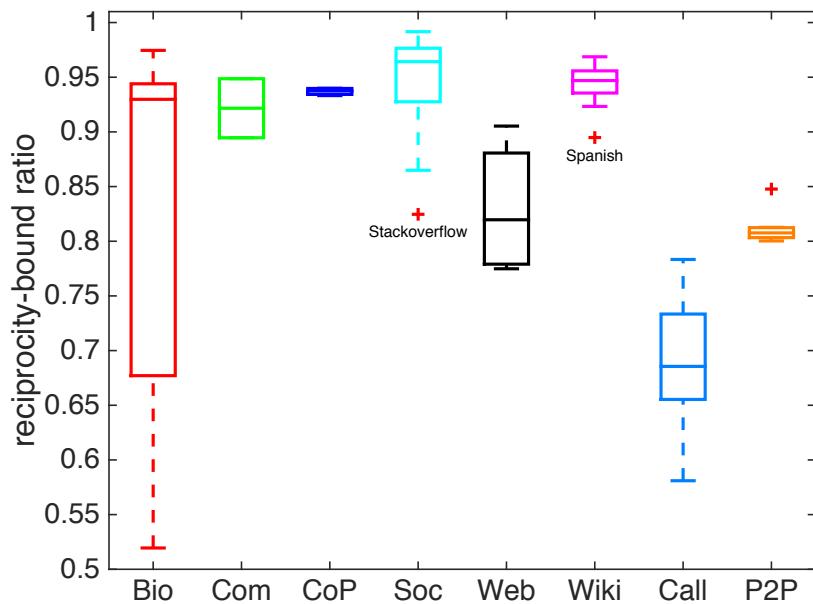


Figure 2.14: Box plot of ratio between reciprocity of 3-path optimal digraph returned by Algorithm 1 and upper bound.

is indeed the case for most of the 3-path optimal digraphs obtained from the real networks studied here, the anti-symmetric parts of which turn out to be acyclic.

2.6 Conclusion

In this chapter, we showed that the maximum reciprocity problem is NP-hard. We provided a partial characterization of networks with maximum reciprocity and a greedy algorithm to eliminate suboptimal motifs. We also provided an upper bound on reciprocity along with necessary conditions and sufficient conditions for achieving the bound. We demonstrated that the bound is surprisingly close to the observed reciprocity in a wide range of real networks, which suggests that the tendency to form reciprocal edges might be much stronger than the observed reciprocity indicates. We found surprising linear relationships between empirical reciprocities and the corresponding upper bounds. We showed that a particular type of suboptimal motif called 3-paths is the major source of loss in reciprocity in these networks.

CHAPTER 3

COMPETITION UNDER CUMULATIVE ADVANTAGE

3.1 Introduction

Growth is a fundamental aspect inherent to most networks that has been widely investigated both empirically through the analysis of data from various contexts and theoretically through idealized models. An important driving force behind network growth and in particular the evolution of node degrees is *cumulative advantage* (CA), where accumulated edges (i.e., current node degree) promote gathering even more edges; see [62] and the reference therein. A second widely-accepted driving force in this context is *fitness*, which captures the inherent ability of nodes to attract edges. Thus, dynamics of network growth is governed by skill (fitness) and luck (random but biased edge attachment).

Recent work has framed the problem of network growth as a competition among nodes [50, 72, 63]. In essence, nodes in a network *compete* with one another to accumulate edges, increasing (or decreasing) their degrees over time. As expected, such competitions are also driven by skill and luck and have been studied empirically and theoretically for different networks, an example of which is the evolution and predictability of success in citation networks [72]. Outside the domain of networks, the study of skill and luck competitions has a long history in social and physical sciences [7, 22].

However, the intricacies of skill and luck competitions are far from trivial, even in a simple CA model with just two competitors. To illustrate, consider a network with two hub nodes that compete for connectivity. Each time a new node joins the

network, it connects to one of the hubs randomly with bias depending on the hubs' fitnesses (model details given in Section 3.3). In the presence of CA, the bias also depends on the hubs' current degrees. Figure 3.1(a) and Figure 3.1(b) illustrate the difference of the hubs' degrees over time for two sample paths in the absence and presence of CA, respectively. The paths with the same color and label in both plots are generated using the same pseudorandom sequence. The competition is tied every time the degree difference is zero and we define the competition duration as the time until the final tie occurs. Note that the red path in Figure 3.1(b) lasts much shorter than the red path in Figure 3.1(a), while the blue path in Figure 3.1(b) lasts significantly longer than the blue path in Figure 3.1(a). This suggests a potentially larger variance in competition duration in the presence of CA. Moreover, with CA the less fit hub may enjoy a sizable degree leadership for a long time; see the blue path in Figure 3.1(b). These observations also apply to two specific nodes in more general network growth models, provided that we interpret "time" as the increase in the total degrees of the two given nodes. However, are these sample paths anomalies or the norm? Can we be more precise about these observations?

In this chapter, we aim to develop a fundamental understanding of the effects of CA in such growth competitions. We approach this problem by considering classical, simple and well-studied theoretical models for competitions based on skill and luck that are either coupled with or free of cumulative advantage. These models may not be general enough as statistical models that fit real-world data for competitions in growing networks, as such models must capture intricate features of the domain, such as skill distribution or amplitude of cumulative advantage (e.g., linear or sub-linear), as well as their time dependency. However, they still provide invaluable insights into how CA impacts competitions. More specifically, we focus on competitions between two agents (nodes) and study two fundamental aspects of competitions: *duration* – the time required for the most skilled to overtake its competitor and forever en-

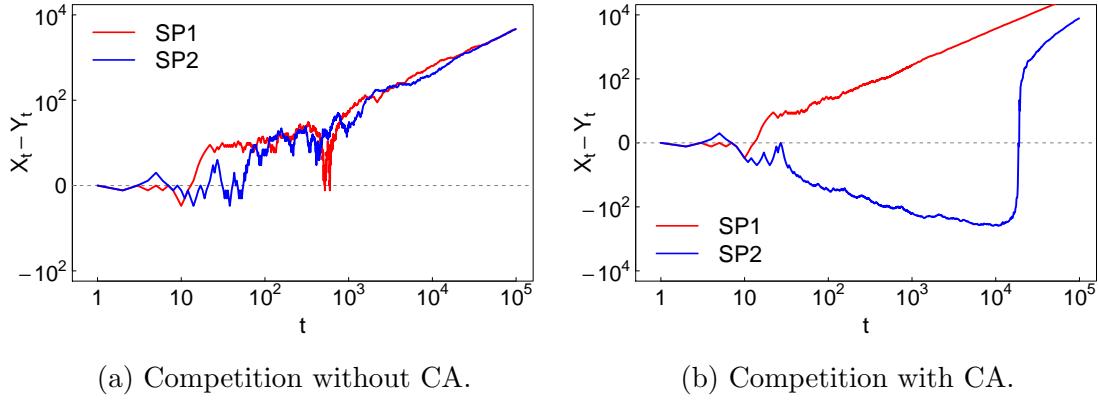


Figure 3.1: Time evolution of degree difference between two nodes in competitions without and with cumulative advantage (CA). Each plot shows two independently simulated sample paths. The sample paths with the same color in both plots use the same sequence of pseudorandom numbers. Competition starts tied and node X is 10% fitter than node Y . See model details in Section 3.3.

joy undisputed leadership; *intensity* – the number of times competitors tie for the leadership. In this direction, we make the following main contributions.

- In the case where the two competitors have equal fitness, we obtain the asymptotic tail distributions for both duration and intensity of CA competitions. We demonstrate that they are power laws with respective tail exponents $-1/2$ and -1 , which are independent of the initial wealth of the competitors.
- In the case where the two competitors have unequal fitness, we derive asymptotic lower and upper bounds for the tail distribution of duration of CA competitions, and an upper bound for the tail distribution of their intensity. These bounds show that duration is heavy tailed while intensity is exponential tailed in the presence of CA. In particular, duration is heavier tailed while intensity is lighter tailed than corresponding RW competitions.
- We observe that a slight difference in fitness of the two results in a extremely heavy tail for duration of CA competitions. Thus, an individual that is only

slightly more skilled than his competitor might have to hang on to the competition for an extremely long period of time before taking the ultimate lead, a phenomenon we call the “struggle of the fittest”.

Despite 90 years since the basic CA model was first proposed [25], known as Pólya’s urn model, the work in this chapter is, to the best of our knowledge, the first to characterize the duration and intensity distributions of CA competitions with skill. We believe our findings have profound implications to our understanding of competitions, beyond its importance to the evolution of degree of nodes in growing networks.

The rest of this chapter is organized as follows. Section 3.2 briefly discusses the related work. Section 3.3 introduces the CA competition model. Section 3.4 presents the theoretical results, illustrated and supplemented by simulations. Section 3.5 concludes this chapter.

3.2 Related Work

Resource accumulation is an ubiquitous phenomenon that naturally arises in a variety of social and complex systems. The problem is usually framed as a competition among agents for resources that are abundant, and has been studied in different contexts across various disciplines ranging from protein binding within a cell [26, 46] to views of online social media [14, 29] and citations among scholarly papers [65, 72]. In the context of networks, network growth and in particular node degree evolution has been recently framed as competition among nodes, where different aspects of competitions have been studied empirically and theoretically [72, 63, 33, 50, 13].

Models for resource accumulation competitions generally incorporate skill (fitness), luck (randomness) and externalities. Cumulative advantage is one type of externality that is considered a general mechanism for inequality [22]. It appears in the literature under many variants such as Price’s cumulative advantage model [65],

preferential attachment [9, 10, 12], “the rich get richer”, Matthew effect [22, 55, 62], and path-dependent increasing returns [7]. The Pólya’s urn model [25, 54] is widely used to capture these effects. Most previous work on Pólya’s urn model and its generalizations focuses on the share of resources gathered by each agent, also known as the agent’s *market share*, proving convergence and limiting results of the market share distribution [54, 43, 61]. More recent studies consider Pólya’s urn models with non-linear bias [23], the effects of initial conditions of urns [59, 15], as well as the time for the first tie [6] and probability of a tie ever occurring [71].

However, two fundamental metrics associated with competitions, *duration* - how long it takes for the undisputed winner to emerge, and *intensity* - how many times the competitors tie for the leadership, have largely been neglected in the literature. Previous results establish that the most skilled agent eventually wins [54], and that average intensity up to time t is approximately $(\log t)^\alpha$, where α depends on the relative skill of the competitors [34, 33]. To the best of our knowledge, no previous work has provided rigorous characterizations for the distributions of duration and intensity of competitions in Pólya’s urn models. This chapter partially fills this gap for the two competitor case and sheds light on some recent approximate results [34, 33].

3.3 Models

In this section, we formally introduce competition models for two competing agents and give precise definitions for two fundamental metrics of a competition, i.e. its duration and intensity.

3.3.1 General Setup and Metrics

Let X and Y denote the two agents that engage in the competition. Each agent is associated with a positive *fitness* value that reflects its intrinsic competitiveness or skill level. Let f_X and f_Y denote the fitness of X and Y , respectively, and $r = f_X/f_Y$

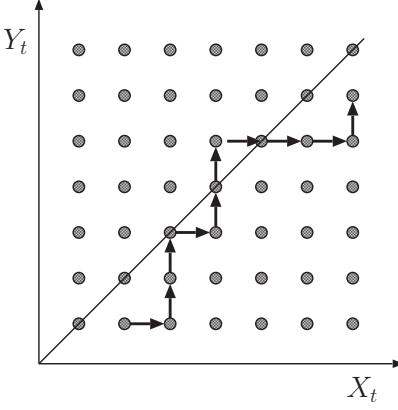


Figure 3.2: State space of competition processes, with an illustration of a sample path with $x_0 = 2$, $y_0 = 1$, and three ties at time $t = 3, 5, 7$.

the fitness ratio. Without loss of generality, we assume that $f_X \geq f_Y$ and hence $r \geq 1$.

The resource that the agents compete for will be generically referred to as wealth, which is measured in discrete units. The competition starts at time $t = 0$ with agents X and Y having x_0 and y_0 units of initial wealth, respectively. We consider a discrete-time process. At each time step, one unit of wealth is added to the system and given to either X or Y . Denote by X_t and Y_t the respective cumulative wealth of X and Y at time t . The complete history of the competition $\{(X_t, Y_t)\}_{t=0}^{\infty}$ then forms a discrete-time discrete-space stochastic process. The state space S is the first quadrant of the integral lattice (see Figure 3.2),

$$S = \{(x, y) \in \mathbb{Z}^2 : x \geq 1, y \geq 1\}.$$

The initial condition is $(X_0, Y_0) = (x_0, y_0)$. How the process evolves over time is determined by specific competition models, of which the CA competition model to be introduced in Section 3.3.2 is an example.

We now make the notions of duration and intensity of competitions more precise by defining them through events of wealth ties. Given a competition process

$\{(X_t, Y_t)\}_{t=0}^{\infty}$, we say that a *tie* occurs at time t if $X_t = Y_t$. Figure 3.2 shows three ties at times $t = 3, 5, 7$.

The *duration* T of a competition is defined to be the time of the last tie, i.e.,

$$T = \sup\{t \geq 0 : X_t = Y_t\}.$$

When there is no tie, we follow the standard convention that $T = \sup \emptyset = -\infty$. The competition ends at time T in the sense that one of the agents takes the lead and never loses it again after T .

The *intensity* N_t of a competition until time t is the number of ties that occur by time t , i.e.,

$$N_t = \sum_{i=0}^t \mathbb{1}_{X_i=Y_i},$$

where $\mathbb{1}_A$ is the indicator of event A . The *intensity* N of a competition is the total number of ties throughout the competition, i.e., $N = \lim_{t \rightarrow \infty} N_t$. This measures the intensity of the competition in the sense that it counts the number of potential changes in leadership. Note that $T < +\infty$ if and only if $N < +\infty$.

3.3.2 CA Competition Model

In the CA competition model, the unit of wealth introduced at time $t+1$ is given to X with probability

$$p_{X,t} = \frac{f_X X_t}{f_X X_t + f_Y Y_t} = \frac{r X_t}{r X_t + Y_t};$$

otherwise it is given to Y . Note that the transition probability $p_{X,t}$ embodies both fitness and CA effects (externalities).

More formally, in the CA competition model, the complete history $\{(X_t, Y_t)\}_{t=0}^\infty$ forms a discrete-time Markov chain with stationary transition probabilities. The transition probability $\mathbb{P}[(X_{t+1}, Y_{t+1}) = (x', y') \mid (X_t, Y_t) = (x, y)]$ is given by

$$Q_{\text{CA},r}(x, y; x', y') = \begin{cases} \frac{rx}{rx + y}, & \text{if } (x', y') = (x + 1, y), \\ \frac{y}{rx + y}, & \text{if } (x', y') = (x, y + 1), \\ 0, & \text{otherwise.} \end{cases} \quad (3.1)$$

Note that the transition probabilities are spatially inhomogeneous, i.e., they depend on the current state (x, y) , which makes the analysis difficult, especially when $r > 1$.

For the purpose of comparison, a RW competition model incorporates skill and luck but not the CA effect (no externalities), where the transition probabilities are determined entirely by the fitness ratio r . In particular, the probability that agent X receives the unit of wealth introduced at any time is always given by

$$p_X = \frac{f_X}{f_X + f_Y} = \frac{r}{r + 1}.$$

Thus the RW competition model is a discrete-time Markov chain with the same state space S as the CA competition model, but with the following spatially homogeneous transition probabilities,

$$Q_{\text{RW},r}(x, y; x', y') = \begin{cases} \frac{r}{r + 1}, & \text{if } (x', y') = (x + 1, y), \\ \frac{1}{r + 1}, & \text{if } (x', y') = (x, y + 1), \\ 0, & \text{otherwise.} \end{cases}$$

The spatial homogeneity of the transition probabilities leads to a more tractable analysis. In fact, the difference process $\{X_t - Y_t\}$ is a standard biased RW with parameter $r/(r + 1)$. Thus the abundance of known results for RW [38] can be

directly translated into results for RW competitions, including duration and intensity as we have defined in Section 3.3.1.

Throughout the rest of this chapter, we use CA_\equiv and RW_\equiv to denote CA and RW competitions with identical fitness ($r = 1$), respectively. We use CA_\neq and RW_\neq to denote CA and RW competitions with distinct fitnesses ($r > 1$), respectively. Before presenting our results, we point out some connections between the CA and RW models that are useful in our analysis. In particular, in CA_\equiv , all paths connecting two given states (x_0, y_0) and (x, y) have the same probability. This is a nice property that CA_\equiv shares with RW, which enables us to leverage existing results on RW in our analysis of CA_\equiv . Unfortunately, this property is lost in CA_\neq , where we resort to the Chapman-Kolmogorov equation for upper and lower bounds on the probabilities of interest. In the limiting case where X_t and Y_t are both large but comparable to each other, the connection to RW is again partially retained, a fact we also exploit in the analysis of CA_\neq .

3.4 Results

In this section we present our theoretical results for duration and intensity distributions, which are also illustrated graphically and supported by extensive numerical simulations. Table 3.1 provides a summary of our main results along with prior knowledge about RW competitions from the literature. Note that $\mathbb{P}_{\langle \text{MODEL} \rangle, r}^{(x_0, y_0)}$ denotes the probability in model $\langle \text{MODEL} \rangle \in \{\text{CA}, \text{RW}\}$ with fitness ratio r and initial state (x_0, y_0) . The following notations have been used in Table 3.1 and will be used throughout the rest of this chapter.

- $f(x) \sim g(x)$ if and only if $\lim_{x \rightarrow \infty} f(x)/g(x) = 1$.
- $f(x) \lesssim g(x)$ if and only if $\limsup_{x \rightarrow \infty} f(x)/g(x) \leq 1$.
- $f(x) \gtrsim g(x)$ if and only if $\liminf_{x \rightarrow \infty} f(x)/g(x) \geq 1$.

metric	$\langle \text{MODEL} \rangle$	$r = 1$	$r > 1$
duration T : $\mathbb{P}_{\langle \text{MODEL} \rangle, r}^{(x_0, y_0)}[T \geq t]$	CA	$\sim t^{-1/2}$	$\lesssim t^{-(r-1)x_0}$ $\gtrsim t^{-(r-1)(x_0 - \frac{1}{r})}$
	RW	1	$\leq \left[\frac{4r}{(r+1)^2} \right]^t$
intensity N : $\mathbb{P}_{\langle \text{MODEL} \rangle, r}^{(x_0, y_0)}[N \geq n]$	CA	$\sim n^{-1}$	$\leq \left(\frac{2}{r+1} \right)^{n-1}$
	RW	1	$\left(\frac{2}{r+1} \right)^{n-1}$

Table 3.1: Tail distributions for duration and intensity of competitions in both RW and CA models. Multiplicative constants are omitted in all expressions involving t and n . The RW statistics can be found in most textbooks on the topic, e.g. [38, pp. 113,116].

All proofs are relegated to Appendix C.

3.4.1 Competition Duration

As shown in Table 3.1, RW_\equiv competitions never end, i.e., $\mathbb{P}_{\text{RW}, 1}^{(x_0, y_0)}[T = \infty] = 1$, while RW_\neq competitions are generally very short, whose durations exhibit exponential tails. The story for CA competitions is drastically different. The introduction of CA guarantees that a competition always ends, i.e. $\mathbb{P}_{\text{CA}, r}^{(x_0, y_0)}[T < \infty] = 1$, even when the two agents are equally fit, which is in sharp contrast to endless RW_\equiv competitions. On the other hand, CA fundamentally *increases* the chance of having a long-lasting competition between unequally fit agents, as the duration of CA_\neq always has a power-law distribution, in contrast to a sub-exponential distribution for RW_\neq . Thus, cumulative advantage does *not* always make competitions shorter as one might expect.

3.4.1.1 Equal Fitness Case: CA_\equiv

The following theorem shows that the duration T for CA_\equiv is heavy-tailed with an asymptotic power-law distribution.

Theorem 3.4.1. *The duration of a CA₊ competition has the following asymptotic tail distribution,*

$$\mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[T \geq t] \sim \frac{1}{2^{x_0+y_0-5/2}\sqrt{\pi}B(x_0,y_0)}t^{-1/2}, \quad (3.2)$$

where $B(x, y) = \int_0^1 s^{x-1}(1-s)^{y-1}ds$ is the beta function.

It follows from (3.2) that

$$\mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[T < \infty] = 1 - \lim_{t \rightarrow \infty} \mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[T \geq t] = 1,$$

i.e. the duration of CA₊ is almost surely finite.

Note, however, that the power-law exponent is always $-1/2$, independent of the initial wealth x_0 and y_0 . Consequently, although the duration of CA₊ is finite rather than infinite as in RW₊, the expected duration is still infinite, even if x_0 is significantly larger than y_0 or vice versa.

On the other hand, the initial wealth (x_0, y_0) does affect the multiplicative factor in (3.2). Figure 3.3 shows the duration distributions from simulations for various values of initial wealth, with the asymptotes in Eq. (3.2) superimposed. Each simulation curve is the average of 10^5 independent runs for 10^7 time steps each. All curves are truncated at $t = 10^6$, since the empirical distributions will drop down sharply and become inaccurate as t approaches the cutoff time in simulations. Similar truncations will be applied to later plots without further mention. Note the good agreement between theory and simulation in the tails in Figure 3.3. When both x_0 and y_0 increase but are kept equal, the distribution curve shifts upwards, which means the competition lasts longer. When the initial wealth of only one agent (y_0 here) increases, the distribution curve shifts downwards, which means the competition is shorter.

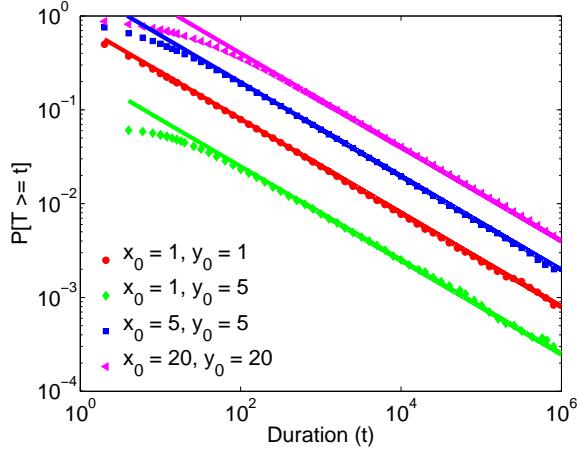


Figure 3.3: Tail distribution for duration of CA_\equiv with various (x_0, y_0) . The dots are simulation results. The solid lines are the asymptotes in Eq. (3.2).

3.4.1.2 Different Fitness Case: CA_\neq

The next theorem shows that the tail distribution of the duration T for CA_\neq is asymptotically bounded by power laws from both above and below.

Theorem 3.4.2. *The tail distribution of the duration of a CA_\neq competition has the following asymptotic bounds,*

$$\varphi_1 t^{-(r-1)x_0} \lesssim \mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[T \geq t] \lesssim \varphi_2 t^{-(r-1)(x_0-1/r)}, \quad (3.3)$$

where

$$\varphi_1 = \frac{\Gamma(rx_0 + y_0)}{(r+1)x_0 2^{x_0+y_0-1} \Gamma(x_0) \Gamma(y_0)}, \quad (3.4)$$

and

$$\varphi_2 = \frac{2^{(r-1)(x_0-r^{-1})} \Gamma(r^{-1}) \Gamma(rx_0 + y_0)}{(r+1)(x_0 - r^{-1}) \Gamma(x_0) \Gamma(y_0)}, \quad (3.5)$$

where $\Gamma(x) = \int_0^\infty s^{x-1} e^{-s} ds$ is the gamma function.

It follows from the lower bound that $\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[T < \infty] = 1$ for $r > 1$, i.e. the duration for CA_\neq is almost surely finite as is for CA_\equiv . The constants φ_1 and φ_2 are very loose, so the bounds are best interpreted as bounds on the tail exponent.

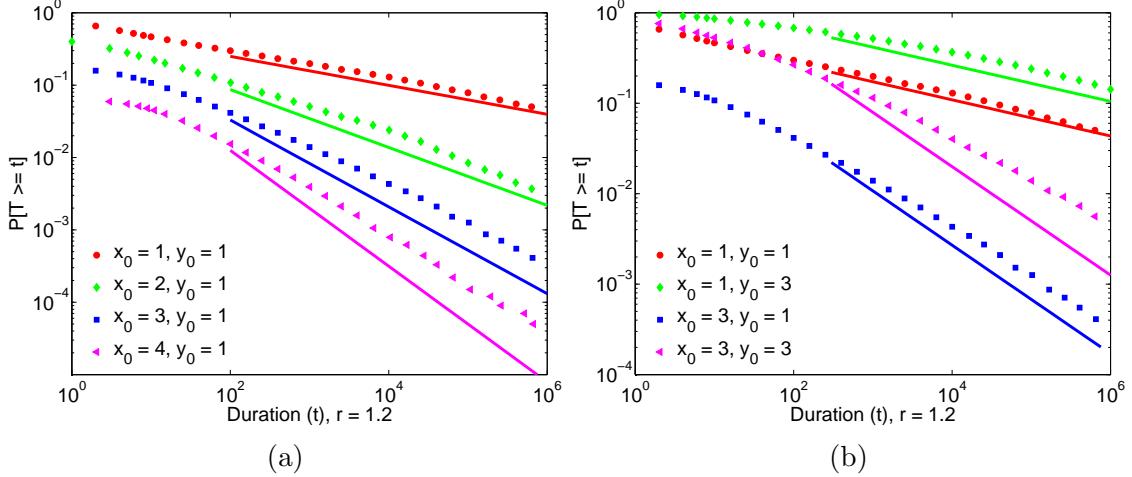


Figure 3.4: Tail distribution for duration of CA_\neq with $r = 1.2$ and various (x_0, y_0) . The dots are simulation results. The solid lines are the asymptotic lower bound in Eq. (3.3) but shifted closer to the simulation results for easier visual comparison of the slopes.

Note that the power-law exponents in the upper and lower bounds depend on x_0 but not on y_0 , and they differ only by $1 - 1/r < 1$. In this sense, the shape of the distribution at large t is largely determined by the fitness ratio and the initial wealth of the fitter agent, while the initial wealth of the less fit plays a much weaker role. This is illustrated in Figure 3.4, which shows the duration distributions from simulations for $r = 1.2$ and various values of (x_0, y_0) , alongside the lower bounds from Eq. (3.3) that are shifted closer to the simulation results for easier comparison of the slopes. Each simulation curve is the average of 10^5 independent runs for 10^9 time steps each.

Figure 3.4(a) shows how the slopes of the distribution curves, which correspond to the power-law exponents, depend critically on x_0 . The impact of x_0 is two-fold. As x_0 increases, the distribution curve becomes more tilted as predicted by the bounds. At the same time, it also shifts downwards. Both changes mean that the competition tends to be shorter.

Figure 3.4(b) shows the impact of changing both x_0 and y_0 . When x_0 is fixed, increasing y_0 only results in a slight decrease in the absolute value of the slope, in agreement with Eq. (3.3). The distribution curve shifts upwards, which means the competition tends to last longer. When both x_0 and y_0 increase, the situation becomes more intricate. The curve may shift upwards while bending down faster in the tail, which could possibly lead to a crossover in the old and new curves, as is the case of going from $(x_0, y_0) = (1, 1)$ to $(x_0, y_0) = (3, 3)$. In this case, the new competition is more likely to have a medium duration.

3.4.1.3 Struggle-of-the-Fittest Phenomenon

Now we look at the impact of fitness ratio r on duration. Contrasting Eqs. (3.2) and (3.3) leads to an interesting observation. Departing from CA_\equiv by slightly increasing the fitness ratio r from 1 to $1 + \varepsilon$, where ε is close to 0, precipitates a *significant increase* in the probability of long-lasting competitions, as manifested in the discontinuous jump in the power-law exponents from $-1/2$ in Eq. (3.2) to $-\varepsilon x_0 \approx 0$ in Eq. (3.3). This is opposite to what happens in RW competitions, where a slight increase in fitness departing from RW_\equiv to RW_\neq transforms the competition from one that never ends to one with a geometrically distributed duration. The lower bound in Eq. (3.3) shows that CA_\neq with $r < 1 + (2x_0)^{-1}$ is more likely to have long-lasting competitions than CA_\equiv , despite the fact that the fitter agent is bound to become the ultimate winner. We refer to the phenomenon that the fitter agent takes an extremely long time to win as “struggle of the fittest”.

Figure 3.5 shows the duration of simulated CA competitions for various fitness ratios r . Each simulation curve is the average of 10^5 independent runs for 10^9 time steps each. Note how the distribution of duration jumps upward from the curve for CA_\equiv to the curve for CA_\neq with $r = 1.1$. It also shows how the curves for CA_\neq

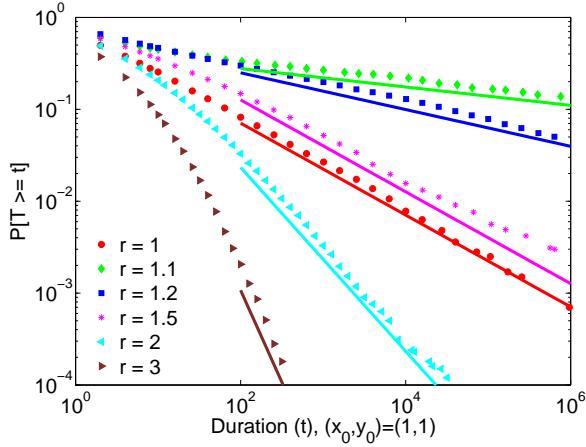


Figure 3.5: Tail distribution for duration of CA with various r . The dots are simulation results. For $r = 1$, the solid line is the asymptote in Eq. (3.2). For $r > 1$, the solid lines are the lower bound in Eq. (3.3) but shifted as in Figure 3.4.

become more and more tilted as r increases, being roughly parallel to the CA_+ curve at $r = 1 + (2x_0)^{-1} = 1.5$.

3.4.2 Competition Intensity

Given that CA competitions are long-lasting, one might expect them also to be intense, i.e., exhibit many ties ($X_t = Y_t$). As we will see in this section, this intuition is appropriate for CA_+ but not for CA_\neq .

3.4.2.1 Equal Fitness Case: CA_+

The following theorem shows that the intensity N of CA_+ is heavy-tailed with an asymptotic power-law distribution.

Theorem 3.4.3. *The intensity of a CA_+ competition has the following asymptotic tail distribution,*

$$\mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[N \geq n] \sim \frac{1}{2^{x_0+y_0-2} B(x_0, y_0)} n^{-1}, \quad (3.6)$$

where $B(x_0, y_0)$ is the beta function as in Eq. (3.2).

In this case, the intensity has infinite expectation, as does the duration. Figure 3.6 shows the duration distributions from simulations for various values of initial wealth,

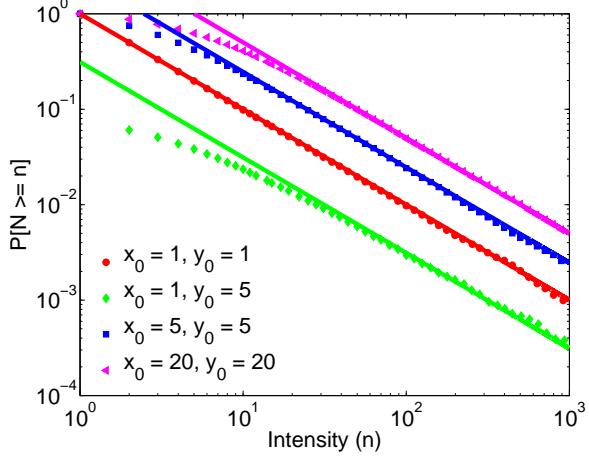


Figure 3.6: Tail distribution for intensity of CA_\equiv with various (x_0, y_0) . The dots are simulation results. The solid lines are the asymptotes from Eq. (3.6).

with the asymptotes in Eq. (3.6) superimposed. Each simulation curve is the average of 10^5 independent runs for 10^7 time steps each. We observe the same behavior as in Figure 3.3. When both x_0 and y_0 increase but are kept equal, the distribution curve shifts upwards, which means the competition is more intense. When the initial wealth of only one agent (y_0 here) increases, the distribution curve shifts downwards, which means the competition is less intense.

We mention in passing that if we have a finite observation time t_f , the expected intensity N_{t_f} by time t_f grows as $\log t_f$, a phenomenon observed for the related CA model in Godrèche et al. [33].

3.4.2.2 Different Fitness Case: CA_\neq

In sharp contrast, CA_\neq competitions are not intense despite their long durations. In fact their intensities are surprisingly mild, bounded above by a geometric distribution, as shown in the next theorem.

Theorem 3.4.4. *The tail distribution of the intensity of a CA_\neq competition has the following upper bound,*

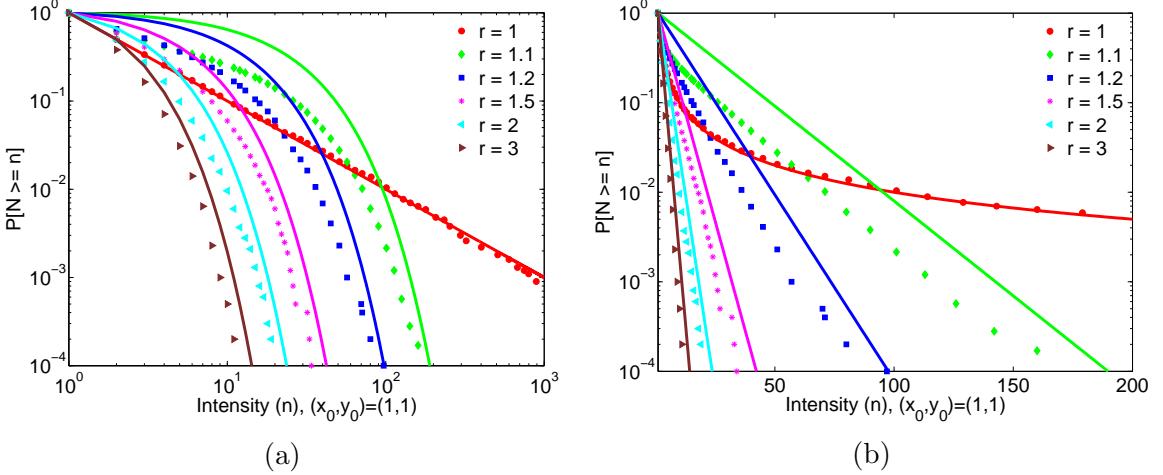


Figure 3.7: Tail distribution for intensity of CA with various r . The dots are simulation results. The solid lines are the upper bounds from Eq. (3.7) for $r > 1$, and the asymptote in Eq. (3.6) for $r = 1$.

$$\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[N \geq n] \leq C \left(\frac{2}{1+r} \right)^{n-1}, \quad (3.7)$$

with

$$C = \begin{cases} 1, & x_0 \leq y_0, \\ \frac{(y_0)_{x_0-y_0}}{(rx_0+y_0)_{x_0-y_0}} \left(1 + \frac{1}{r}\right)^{x_0-y_0}, & x_0 > y_0, \end{cases}$$

where $(x)_k = \prod_{i=0}^{k-1} (x+i)$ is the Pochhammer symbol.

Note that the expectation and all higher moments of N are finite. Therefore, the intensity of a CA competition changes dramatically when the fitnesses of the two parties become unequal, the distribution shifting from a power-law tail to an exponential tail. This is illustrated in Figure 3.7, where each simulation curve is the average of 10^5 independent runs for 10^9 time steps each. An important observation is that both CA_\neq and RW_\neq competitions have intensities that are upper bounded by identical exponential tails (see Table 3.1), while exhibiting fundamentally different durations.

Why are CA_{\neq} competitions simultaneously not intense and long-lasting? The answer resides in the probability of Y being the eventual winner. In $\text{CA}_=$ competitions, Y wins with probability $y_0/(x_0 + y_0)$, while in CA_{\neq} competitions Y (the less fit) never wins. However, for small values of r , especially for those very close to one, the dynamics in the initial stages of the competition closely follows that of $\text{CA}_=$. Thus there is a non-negligible chance that Y takes a significant lead, with the CA effect helping it uphold the lead for a long period of time over which there is no tie. Eventually, however, the fitness effect outweighs the CA effect, and X catches up with Y . By then they both have large accumulated wealth, which makes CA_{\neq} behave like RW_{\neq} in the vicinity of $X = Y$, allowing X to quickly establish a lead ahead of Y . At this final stage both fitness and CA effects work in favor of X , and Y stands little chance in taking the lead again. To summarize, the less fit agent has a non-negligible probability of taking an early lead which can last for a very long time due to the CA effect, but it will ultimately surrender the lead to the fitter agent and never lead again, a phenomenon that we call “delusion of the weakest”, which is the flip-side of “struggle of the fittest”.

Figure 3.8 illustrates this observation by showing sample paths for different values of r , all generated using the same sequence of random numbers. Note that for $r = 1$ (identical fitness), Y wins quickly, whereas for $r = 1.1$ the fitter agent X , having trailed behind for a long time, eventually takes over after 69,426 time steps. Finally, for both $r = 1.2$ and $r = 1.5$ agent X has no trouble quickly winning the competition. These sample paths showcase the long struggle of the “slightly” fitter agent in competitions with CA effects.

3.4.3 Interplay of Duration and Intensity

In this section, we study the relationship between duration and intensity. Note that duration gives a natural upper bound $N \leq T/2$ for intensity, i.e., the number

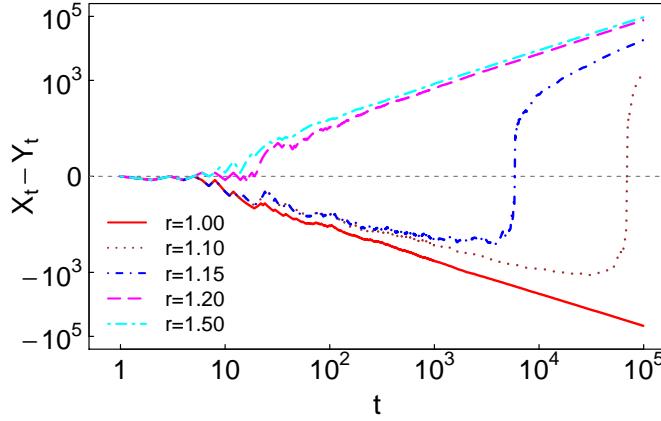


Figure 3.8: “Delusion of the weakest”: sample paths for different values of r ($x_0 = y_0 = 1$), all generated using the same sequence of random bits.

of ties is at most half of the duration in any competition. In CA_\equiv , duration and intensity are strongly and positively correlated. In fact, a tie at time t increases the probability of having another tie at a time greater than t . More precisely, [6] shows that for CA_\equiv ,

$$\mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[T > t | X_t = Y_t] \simeq 1 - \frac{1}{\sqrt{\pi X_t}}. \quad (3.8)$$

Since $X_t \sim t/2$ at a tie, Eq. (3.8) implies that the later a tie occurs, the more likely another tie will occur, intuitively explaining why long-lasting competitions are also intense in this case.

Figure 3.9 shows a scatter-plot of duration versus intensity from 10^4 independent runs of CA_\equiv competitions with $x_0 = y_0 = 1$, each simulated for 10^9 time steps. This unveils a strong positive linear correlation between the two statistics in log-log scale (sample Pearson correlation coefficient of 0.94).

Interestingly, CA_\neq shows a different behavior, since even long-lasting competitions exhibit only a small number of ties. Figure 3.10 shows simulation results for conditional average intensities of competitions with $x_0 = y_0 = 1$ and different fitness ratios r , conditioned on the duration being at least t . Each simulation curve is obtained from 10^4 independent runs for 10^9 time steps each. Note that for $r = 1$, the

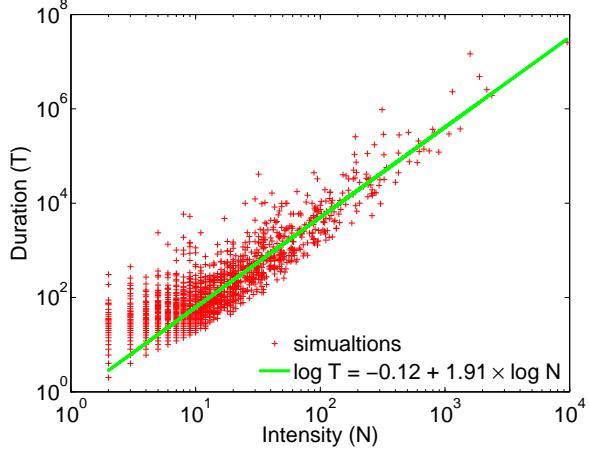


Figure 3.9: Scatter plot of duration vs. intensity for CA₌.

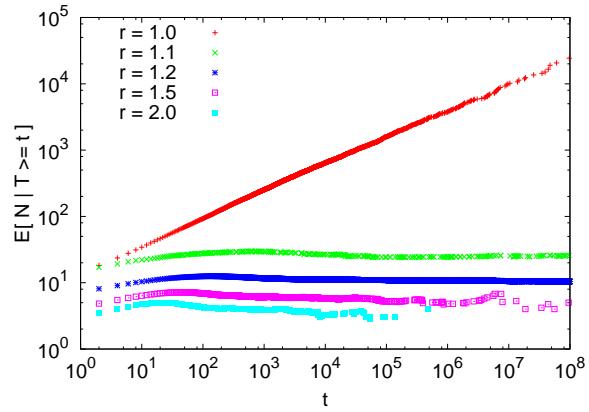


Figure 3.10: Conditional average intensity of competitions conditioned on their duration being at least t , namely $\mathbb{E}_{\text{CA},r}^{(x_0,y_0)}[N \mid T \geq t]$.

conditional average intensity increases linearly with t , but for $r > 1$, it stabilizes as t increases. Again, we observe a sharp transition as we move from identical to distinct fitnesses, this time in the correlation between intensity and duration.

3.5 Discussion and Conclusion

As various empirical studies [67, 72, 69, 22, 63, 62] suggest that real world competitions for resource accumulation are subject to cumulative advantage effects, at least to some extent, a theoretical understanding of the role of skill and luck in com-

petition dynamics becomes a pressing issue. Recent theoretical studies [21, 6, 50, 33] have contributed in this direction.

However, contrary to prior theoretical works, we considered simple and classical mathematical models that capture just the essence of skill and luck competitions with and without CA effects, and investigated fundamental aspects of competition, namely duration (i.e., time until ultimate winner emerges) and intensity (i.e., number of ties in competition). By considering simple models and simple properties we proved and illustrated fundamental theoretical results: CA effect exacerbates the role of luck – power-law tail duration emerges regardless of skill differences, and become extreme (i.e., infinite mean) when skill differences are small enough. Moreover, duration is long not necessarily because of intense competition where agents tussle aggressively for ultimate leadership. On the contrary, under CA, competitions are generally very mild, exhibiting an exponential tail. Long competitions emerge when an early stroke of luck places the less skilled in the lead, who can then, boosted by CA effects, enjoy leadership for a very long period of time. Thus, when CA is present luck sides with the less skilled.

The non-negligible probability of long-lasting competitions has far-reaching implications. In the absence of CA, it takes very little time for the fittest agent to establish dominance, so it is often reasonable to neglect the possibility of a premature burnout. Such observations are in hand with the “survival of the fittest” principle, since soon enough the more skilled will prevail. In the presence of CA, however, even agents with superior fitness may face the challenge of having to endure extremely long competitions. This challenge becomes all more real when the fitness superiority is only minimal. Will the more skilled survive the seemingly eternal inferiority during the competition? Under CA time becomes a central issue, with delusion becoming reality if the more skilled burns out during a long struggle. Thus, in the face of CA,

the fittest survives only if it can persist, which prompts us to rename the principle “survival of the fittest and persistent” when considering CA competitions.

This observation may also shed light on the seemingly inherent difficulty of predicting success in real-world competitions by observing ongoing sample paths. Different empirical studies have alluded to this problem [67, 72, 69] as well as recent model driven studies [21]. Since direct competitions tend to be relatively very short and skill-differences tend to be small it is well very possible that the winner when competition ended was not the fittest, adding to the difficulty of making accurate predictions.

CHAPTER 4

INFORMATION DISSEMINATION IN SOCIAL NETWORKS UNDER LIMITED BUDGET OF ATTENTION

4.1 Introduction

Information dissemination has been transformed by the emergence of online social networks and their enthusiastic adoption by users. Users rely on trust relationships in social networks for accessing information. Relationships form on the basis of the quality of information received, and in turn determine the speed of propagation in the network.

The literature on information propagation in social networks, or rumor spreading [64] is wide and varied; we mention here only those that are most relevant to the work in this chapter. Previous work has mostly studied the propagation of a rumor originating at a given source. The typical model is the randomized broadcast model [47] which is carried out in synchronized rounds. In each round, each user selects a neighbor at random and propagates a rumor. The spreading mechanisms considered have been broadly of three types: push mechanisms where the user sends the rumor if he has it, to the chosen neighbor; pull mechanisms where the user pulls the rumor from the chosen neighbor; and a combined mechanism, where the user pushes the rumor if he has it and pulls it if the chosen neighbor has it. Most previous work focus on the characterization of the delay in spreading a single rumor to all nodes (e.g. [20], [17]). An asynchronous model is considered by Amini et al. [4], where each node contacts a neighbor after a random amount of time; they focus in particular on random regular graphs and derive performance results in terms of optimal delay.

What has been overlooked in most previous work is the practical constraint that each user has a limited *budget of attention*, that is, a limited frequency with which he interacts with his neighbors. Each user then has to allocate his limited budget of attention among his neighbors. There is some work on analyzing the allocation patterns using real world data. In particular, Backstrom et al. [8] analyze a data set of real measurements on how users split their attention among their friends on Facebook. They consider activities such as communication and viewing and show how the balance of attention varies across activities and other personal characteristics. Since the level of attention affects information flow [49], the allocation of the budgets of attention will have an impact on information propagation.

This chapter considers a scenario of information propagation where there are multiple information sources in the network and each user has a limited budget of attention. Each user allocates his budget strategically with the objective of timely reception of news. We consider an asynchronous pull model, where each node contacts a neighbor after a random delay and pulls any content available at that neighbor and the frequency with which he pulls content from neighbors is limited. Our objective is to identify optimal allocations of this limited frequency among neighbors for each user in the network and to inform the design of algorithms for optimal information spread. In particular, we want to answer the question “when users make selfish decisions on how to allocate their limited access frequency among neighbors, does information propagate efficiently?” We take the approach of conceiving a general model for studying the balance of attention, and an analysis for several network topologies.

We investigate the efficiency of selfish allocation by considering the metric of average end-to-end delay of content spread. We make the following contributions:

- We study the efficiency of selfish allocation of the budget of attention by characterizing the *price of stability* (PoS) for several network topologies.

- We identify topologies with inefficient propagation under selfish allocations.
- We propose the “plus-one” mechanism, an incentive scheme that coaxes users into mimicking a gradient-descent algorithm, bringing the cost of content propagation closer to the optimal.
- We present numerical results that compare the optimal, selfish, and feedback-based allocations.

The rest of this chapter is organized as follows. We define the social network model in Section 4.2, and present the analysis of optimal and distributed allocation over several topologies in Section 4.3. In Section 4.4 we present our feedback-based mechanism. Numerical results follow in Section 4.5 and we conclude in Section 4.6.

4.2 Model Description

We consider a social network where each user has a set of friends, or *contacts*, that he links to, or *follows*, in order to get content such as news updates, videos, or other messages. Each user then makes all content that he holds available to his followers, such as on the Facebook wall or Twitter stream. Rather than a single source of content, we allow all users to create content, at a given rate. Further, we assume that users seek to obtain all content circulating in the network. Users consult their contacts for the latest updates of information with the objective of minimizing the average delay for obtaining all information. As in real online social networks, users have a limited *budget of attention*, that is, the total rate at which they may consult their contacts is limited. As such, this rate must be allocated among the contacts in a manner that optimizes for delay. We will compare a centralized optimization of consultation rates with a distributed one where users optimize the allocation in a selfish manner.

We model the social network as a directed graph $G = (V, E)$, where V is the set of users, and E is the set of directed edges between users, i.e. $(i, j) \in E$ if and only if there is an edge from i to j . Denote by $N(i)$ the set of in-neighbors of user i , i.e. $N(i) = \{j \in V : (j, i) \in E\}$, and let $d_i = |N(i)|$ be the in-degree of i . We assume that G is strongly connected.

Each user $i \in V$ creates contents according to a Poisson process with rate $\lambda_i > 0$. When the λ_i 's are the same, we denote the common value by λ . User i consults his in-neighbor $j \in N(i)$ according to a Poisson process with rate x_{ji} . Each user has a limited budget of attention $b_i > 0$, so that the rates of consultation are constrained by $\sum_{j \in N(i)} x_{ji} = b_i$. When the b_i 's are the same, we denote the common value by b . In terms of the normalized rates $y_{ji} = x_{ji}/b_i$, the budget constraint becomes

$$\sum_{j \in N(i)} y_{ji} = 1. \quad (4.1)$$

The vector $\mathbf{y}_i = \{y_{ji} : j \in N(i)\}$ represents how user i allocates his budget of attention among his neighbors, which will be referred to as his strategy. The set of all possible strategies of user i is the unit simplex Δ_i in \mathbb{R}^{d_i} . Let $\mathbf{y} = \{\mathbf{y}_i : i \in V\}$ be the strategy profile of the network, and $\Delta = \bigtimes_{i \in V} \Delta_i$ is the set of all possible profiles. Also let \mathbf{y}_{-i} be the strategy profile of all users other than i .

For any $i \neq j$, define $D_{ji}(\mathbf{y})$ to be the delay for user i to receive content originated at user j under profile \mathbf{y} . Define the cost for user i to be

$$C_i(\mathbf{y}) = \frac{1}{\lambda_{-i}} \sum_{j \in V \setminus \{i\}} \lambda_j \mathbb{E} D_{ji}(\mathbf{y}), \quad (4.2)$$

where $\lambda_{-i} = \sum_{j \in V \setminus \{i\}} \lambda_j$, and define the social cost to be

$$C(\mathbf{y}) = \frac{1}{(n-1)\Lambda} \sum_{i \in V} \sum_{j \in V \setminus \{i\}} \lambda_j \mathbb{E} D_{ji}(\mathbf{y}), \quad (4.3)$$

where $\Lambda = \sum_{i \in V} \lambda_i$. Note that both (4.2) and (4.3) are independent of any normalization of the λ_i 's, which just translates a change of units. We set $\lambda = 1$ when content creation rates are homogeneous.

Let $C^* = \min_{\mathbf{y} \in \Delta} C(\mathbf{y})$ be the optimal social cost, and \mathbf{y}^* a profile that minimizes the social cost C . Let $\Gamma \subset \Delta$ be the set of profiles that are Nash equilibria when each user minimizes his own cost in a selfish manner. In general there might exist multiple equilibria. A standard measure of inefficiency of equilibria is the *price of anarchy* (PoA). It is defined as the ratio of the cost at the worst equilibrium to that of an optimal outcome, i.e., $\max_{\mathbf{y} \in \Gamma} C(\mathbf{y})/C^*$. We are, however, interested in the best equilibria, which would give us a benchmark of what's achievable through distributed means. As such, we focus on the *price of stability* (PoS), defined as the ratio of the cost at the best equilibrium to that of an optimal outcome [58], i.e.,

$$\text{PoS} = \min_{\mathbf{y} \in \Gamma} \frac{C(\mathbf{y})}{C^*} = \frac{\widehat{C}}{C^*}, \quad (4.4)$$

where $\widehat{C} = \min_{\mathbf{y} \in \Gamma} C(\mathbf{y}) \geq C^*$ is the minimum social cost under a selfish allocation. Even though the PoS can be seen as a weaker notion of inefficiency, we find it more interesting in a practical sense, since it gives us a target performance for the design of distributed algorithms. In some cases, e.g., in a tree network, the Nash equilibrium is unique, so the PoS coincides with the PoA.

In what follows, we will measure inefficiencies in the selfish allocation of the budget of attention in several network topologies. We will show that some topologies lead to large inefficiencies. In Section 4.4 we propose a feedback-based mechanism that results in a distributed allocation that has a cost closer to the optimal cost than does the selfish allocation.

4.3 Efficiency Analysis

We will now study the efficiency of selfish optimization on several social network topologies. The interest in studying various topologies lies not only in understanding how existing social networks with those topologies perform, but also in identifying efficient structures for information dissemination. The latter may inform smart design for information propagation.

4.3.1 Tree Network

We first consider tree topologies. Such structures are interesting as networks since information dissemination can be locally tree-like. Let G be a tree; for all tree networks we study, we will assume G is undirected. Let T_{ji} denote the component containing j when an edge $(j, i) \in E$ is removed. Let $\lambda_{ji} = \sum_{k \in T_{ji}} \lambda_k$ be the aggregate content creation rate of the nodes in T_{ji} , and $n_{ji} = |T_{ji}|$ the number of nodes in T_{ji} .

For $i \neq j$, let $\mathcal{P}_{j \leadsto i}$ be the unique shortest path from j to i . The average delay for user i to get contents originated in j is then $\mathbb{E}D_{ji}(\mathbf{y}) = \sum_{e \in \mathcal{P}_{j \leadsto i}} x_e^{-1}$, where $x_e = x_{uw}$ for an edge $e = (u, w)$. Thus the cost for user i is

$$C_i(\mathbf{y}) = \frac{1}{\lambda_{-i}} \sum_{j:j \neq i} \lambda_j \sum_{e \in \mathcal{P}_{j \leadsto i}} \frac{1}{x_e} = \frac{1}{\lambda_{-i} b_i} \sum_{k \in N(i)} \frac{\lambda_{ki}}{y_{ki}} + f(\mathbf{y}_{-i}),$$

where $f(\mathbf{y}_{-i})$ represents terms that do not depend on \mathbf{y}_i . Note that $f(\mathbf{y}_{-i})$ can be infinity for some \mathbf{y}_{-i} on the boundary of Δ_{-i} . However, since users are trying to collect all the information, even selfish users have incentives to keep \mathbf{y} in the interior of Δ so that no information pathway is effectively cut off. Therefore, we will assume \mathbf{y} is in the interior of Δ , and consequently $f(\mathbf{y}_{-i})$ is finite.

If user i selfishly minimizes C_i , the unique best rate allocation, irrespective of how others allocate their budgets of attention, is given by

$$\hat{y}_{ji} = \frac{\sqrt{\lambda_{ji}}}{\sum_{k \in N(i)} \sqrt{\lambda_{ki}}}, \quad \text{for } j \in N(i). \quad (4.5)$$

Note that in this case each user has a unique best selfish strategy independent of others' strategies. Thus there is a unique Nash equilibrium and the PoS coincides with the PoA.

The social cost can be written as follows,

$$\begin{aligned} C &= \frac{1}{(n-1)\Lambda} \sum_{i \in V} \sum_{j: i \neq j} \lambda_j \sum_{e \in \mathcal{P}_{j \sim i}} \frac{1}{x_e} \\ &= \frac{1}{(n-1)\Lambda} \sum_{(u,w) \in E} \frac{1}{x_{uw}} \sum_{i \in V} \sum_{j: i \neq j} \lambda_j \mathbb{1}_{(u,w) \in \mathcal{P}_{j \sim i}} \\ &= \frac{1}{(n-1)\Lambda} \sum_{(u,w) \in E} \frac{1}{x_{uw}} \sum_{i \in T_{wu}} \sum_{j \in T_{uw}} \lambda_j \\ &= \frac{1}{(n-1)\Lambda} \sum_{(u,w) \in E} \frac{n_{wu}\lambda_{uw}}{y_{uw}b_w} \\ &= \frac{1}{(n-1)\Lambda} \sum_{i \in V} \frac{1}{b_i} \sum_{k \in N(i)} \frac{n_{ik}\lambda_{ki}}{y_{ki}}. \end{aligned}$$

Thus the social cost under selfish allocation (4.5) is given by

$$\hat{C} = \frac{1}{(n-1)\Lambda} \sum_{i \in V} \frac{1}{b_i} \left(\sum_{j \in N(i)} n_{ij} \sqrt{\lambda_{ji}} \right) \left(\sum_{j \in N(i)} \sqrt{\lambda_{ji}} \right). \quad (4.6)$$

However, the socially optimal rate allocation is given by

$$y_{ji}^* = \frac{\sqrt{n_{ij}\lambda_{ji}}}{\sum_{k \in N(i)} \sqrt{n_{ik}\lambda_{ki}}}, \quad \text{for } j \in N(i), \quad (4.7)$$

with the optimal social cost being

$$C^* = \frac{1}{(n-1)\Lambda} \sum_{i \in V} \frac{1}{b_i} \left(\sum_{j \in N(i)} \sqrt{n_{ij}\lambda_{ji}} \right)^2. \quad (4.8)$$

We will now study in more detail specific tree structures: the line, the k -ary tree, and the chained star networks.

4.3.1.1 Line Network

Suppose G is a line network, with $V = \{1, \dots, n\}$ and $(i, i+1) \in E$ for $i = 1, 2, \dots, n-1$ and $(i, i-1) \in E$ for $i = 2, \dots, n$. Theorem 4.3.1 below gives bounds on the range of C^* and \widehat{C} and an upper bound on the PoS. Note that the upper bound on PoS does not depend on the content creation rates λ_i . When the budgets of attention are homogeneous, we have $\text{PoS} \leq 5$, but $C^* = \Theta(n/b)$.

Theorem 4.3.1. *In a line network of $n \geq 2$ nodes, the optimal social cost and the cost under selfish allocation are bounded as follows,*

$$\frac{n+1}{8b_{\max}} \leq C^* \leq \widehat{C} \leq \frac{n+1}{b_{\min}} \left(\frac{9}{8} + \frac{1}{4n-4} \right), \quad (4.9)$$

and

$$\text{PoS} \leq \frac{5b_{\max}}{b_{\min}}, \quad (4.10)$$

where $b_{\max} = \max_i b_i$ and $b_{\min} = \min_i b_i$.

Proof. See Appendix D.1. □

When the budgets of attention are heterogeneous, the upper bound in (4.10) can become arbitrarily large. In this case, the PoS can be arbitrarily large as well, as shown in Theorem 4.3.2.

Theorem 4.3.2. *The PoS can be arbitrarily large when the budgets of attention are heterogeneous in a line network.*

Proof. See Appendix D.2. □

4.3.1.2 Chained Star Network

Consider k star networks, each with p nodes. The hubs of the stars are chained to form a line network, with a total of $n = pk$ nodes. Such topologies are not uncommon in social networks based on communities. Such structure might correspond to communities focused on given topics or interests, that are then connected to the larger social network. Theorem 4.3.3 shows that while the optimal social cost can be large, the PoS is of order 1 in the homogeneous case.

Theorem 4.3.3. *In a chained star network with n users, homogeneous content creation rates $\lambda = 1$ and homogeneous budgets of attention b , the optimal social cost and cost under selfish allocation satisfy*

$$C^* = \Theta(b^{-1} \max\{p, k\}) = \Omega(b^{-1} \sqrt{n}),$$

$$\widehat{C} = \Theta(b^{-1} \max\{p, k\}),$$

with PoS = $\Theta(1)$.

Proof. See Appendix D.3. □

Remark. Suppose $b = 1$. As p changes from $\Theta(1)$ to $\Theta(n)$, C^* can have any order between $\Theta(\sqrt{n})$ and $\Theta(n)$. In particular, $C^* = \Theta(n)$ for $p = 1$ and $p = n$, which correspond to the line and star networks, respectively.

4.3.1.3 k -ary Tree Network

We now consider rooted trees where each node has k children. Such structures are of interest for social networks with few edges, as sparse random graphs are locally tree-like. Theorem 4.3.4 below states the PoS for k -ary trees. Corollary 4.3.5 shows that this PoS can be arbitrarily large for k of constant order and as k scales sublinearly with n , even if the content creation rates and budgets of attention are homogeneous.

Theorem 4.3.4. *In a k -ary complete tree with n users, homogeneous content creation rates $\lambda = 1$ and homogeneous budgets of attention b , the optimal social cost and the cost under selfish allocation satisfy*

$$C^* = \Theta(b^{-1}k \log_k n),$$

$$\widehat{C} = \Theta(b^{-1}k \log_k n + b^{-1}\sqrt{n}),$$

with

$$\text{PoS} = \frac{\widehat{C}}{C^*} = \Theta\left(1 + \frac{\sqrt{n} \log k}{k \log n}\right).$$

Proof. See Appendix D.4. □

Corollary 4.3.5. *Let $b = 1$ and $k = \Theta(n^\alpha)$.*

(1). *If $\alpha = 0$, i.e. $k = \Theta(1)$, the costs are $C^* = \Theta(\log n)$, $\widehat{C} = \Theta(\sqrt{n})$, and*

$$\text{PoS} = \Theta(\sqrt{n}/\log n).$$

(2). *If $0 < \alpha < 1/2$, the costs are $C^* = \Theta(n^\alpha)$, $\widehat{C} = \Theta(\sqrt{n})$, and $\text{PoS} = \Theta(n^{1/2-\alpha})$.*

(3). *If $\alpha \geq 1/2$, the costs are $C^* = \Theta(n^\alpha)$, $\widehat{C} = \Theta(n^\alpha)$, and $\text{PoS} = \Theta(1)$.*

The above corollary verifies the intuition that long thin networks are less efficient for information propagation than wide networks. The former type of network thus would require an incentive-based mechanism to make them more efficient. We will propose one such mechanism in Section 4.4.

4.3.2 Clique Networks

We now consider the case where G is a clique. This is, in some sense, the best-case scenario, where users have the widest choice possible in allocating their budget of attention. The analysis is more involved since there are many paths between each source-destination pair, creating a more complicated dependency structure between

the various links. We consider efficiency for the homogeneous case, where $\lambda_i = 1$ and $b_i = b$. We will show that there exists a selfish profile that is asymptotically optimal, and thus the price of stability is bounded and approaches one as the network size increases.

Let us consider the *uniform* strategy for user $i \in V$, where the consultation rate is $y_{ji} = u_{ji} = b/d_i$ for all $j \in N(i)$. Let $\mathbf{u}_i = \{u_{ji} : j \in N(i)\}$ and $\mathbf{u}_{-i} = \{\mathbf{u}_j : j \in V \setminus \{i\}\}$. The profile $\mathbf{u} = \{\mathbf{u}_i : i \in V\}$ is referred to as the *uniform* profile. For a clique network, $u_{ji} = b/(n-1)$ for all $i \neq j$.

Theorem 4.3.6. *For an n -clique with homogenous content creation rates $\lambda = 1$ and budgets of attention b , the social cost C^u of the uniform profile is $C^u = b^{-1}H_{n-1}$, where $H_n = \sum_{k=1}^n k^{-1}$, the n -th harmonic number.*

Proof. We follow the approach in [41]. Let $D_{j(m)}$ be the delay for a content originating from j to reach at least m other users, i.e. $D_{j(m)}$ is the m -th order statistic of $\{D_{ji} : i \in V \setminus \{j\}\}$ for a given j . Let $V_{jk} = D_{j(k)} - D_{j(k-1)}$, with the convention that $D_{j(0)} = 0$. Thus

$$\sum_{i:i \neq j} D_{ji} = \sum_{m=1}^{n-1} D_{j(m)} = \sum_{m=1}^{n-1} \sum_{k=1}^m V_{jk} = \sum_{k=1}^{n-1} (n-k)V_{jk}. \quad (4.11)$$

Note that V_{jk} is exponentially distributed with parameter $\frac{b}{n-1}k(n-k)$. Taking expectations of (4.11) we get:

$$\sum_{i:i \neq j} \mathbb{E}D_{ji} = \sum_{k=1}^{n-1} (n-k)\mathbb{E}V_{jk} = \sum_{k=1}^{n-1} \frac{n-1}{bk} = \frac{n-1}{b} H_{n-1}.$$

Summing over j and dividing by $n(n-1)$, we obtain the social cost $C^u = b^{-1}H_{n-1}$. \square

The next theorem shows that the uniform strategy by a given user is the best response when all other nodes follow the uniform strategy, showing that this is a Nash equilibrium.

Theorem 4.3.7. *For an n -clique with homogenous content creation rates 1 and budgets of attention b , the uniform profile is a Nash equilibrium.*

Proof. Suppose users $i = 1, 2, \dots, n-1$ follow the uniform strategy and consider user n . For a given j ,

$$D_{jn}(\mathbf{y}_n, \mathbf{u}_{-n}) = \min_{1 \leq k \leq n-1} \left\{ \frac{1}{y_{kn}} X_{kn} + Y_{jk}^n \right\},$$

where $\{X_{kn}\}_{k=1}^{n-1}$ are i.i.d. exponential random variable with parameter b , and Y_{jk}^n is the time for an item originating from j to reach k without passing through n . Note that $\{X_{kn}\}_{k=1}^{n-1}$ are independent of $\{Y_{jk}^n\}_{k=1}^n$. Let $\mathbf{T}_j^n = \{T_{ji}^n\}_{i=1}^{n-1}$ be the order statistics of $\mathbf{Y}_j^n = \{Y_{jk}^n\}_{k=1}^{n-1}$. By symmetry, the random vectors $(Y_{j1}^n, \dots, Y_{j,n-1}^n)$ and $(T_{j\sigma(1)}^n, \dots, T_{j\sigma(n-1)}^n)$ are identically distributed, where σ is a permutation chosen uniformly randomly from the symmetric group S_{n-1} , independently of \mathbf{T}_j^n . Therefore,

$$\begin{aligned} & \mathbb{P}[D_{jn}(\mathbf{y}_n, \mathbf{u}_{-n}) > x \mid \mathbf{T}_j^n] \\ &= \mathbb{P} \left[\bigcap_{k=1}^{n-1} \{X_{kn} > y_{kn}(x - T_{j\sigma(k)}^n)\} \mid \mathbf{T}_j^n \right] \\ &= \mathbb{E}_\sigma \left[\exp \left\{ -b \sum_{k=1}^{n-1} y_{kn}(x - T_{j\sigma(k)})^+ \right\} \right] \\ &\geq \exp \left\{ -b \sum_{k=1}^{n-1} y_{kn} \mathbb{E}_\sigma(x - T_{j\sigma(k)})^+ \right\} \\ &= \exp \left\{ -b \sum_{k=1}^{n-1} y_{kn} \sum_{l=1}^{n-1} \frac{1}{n-1} (x - T_{jl})^+ \right\} \\ &= \exp \left\{ -b \sum_{l=1}^{n-1} \frac{1}{n-1} (x - T_{jl})^+ \right\} \\ &= \mathbb{P} [D_{jn}(\mathbf{u}_n, \mathbf{u}_{-n}) > x \mid \mathbf{T}_j^n], \end{aligned}$$

where the inequality follows from Jensen's inequality. Therefore, for $j = 1, 2, \dots, n-1$,

$$\mathbb{P} [D_{jn}(\mathbf{y}_n, \mathbf{u}_{-n}) > x] \geq \mathbb{P} [D_{jn}(\mathbf{u}_n, \mathbf{u}_{-n}) > x].$$

Integrating over x from 0 to ∞ , we obtain $\mathbb{E}D_{jn}(\mathbf{y}_n, \mathbf{u}_{-n}) \geq \mathbb{E}D_{jn}(\mathbf{u}_n, \mathbf{u}_{-n})$ and hence $C_n(\mathbf{y}_n, \mathbf{u}_{-n}) \geq C_n(\mathbf{u}_n, \mathbf{u}_{-n})$ for any $\mathbf{y}_n \in \Delta_n$. Therefore, the uniform profile is a Nash equilibrium. \square

The next theorem shows that the cost of the uniform profile is larger than the optimal social cost by at most $1/b$. Thus the price of stability approaches one as the network size increases.

Theorem 4.3.8. *In any network of size n where every user publishes at rate $\lambda = 1$ and has a budget of attention b , the optimal social cost for any user i is lower bounded by*

$$C^* \geq \frac{n}{b(n-1)} H_{n-1} - \frac{1}{b} \geq C^u - \frac{1}{b}.$$

Proof. The argument uses a *backwards growth process* similar to that used in Section V of [39]. Assume the process is in steady state, i.e. the process started at $-\infty$. Consider user i . Let $B_i(t)$ be the set of users whose states at time $-t$ reach user i by time 0. Note that $B_i(0) = \{i\}$. By stationarity and the independence of the publishing and consulting processes, $\mathbb{P}[D_{ji} > t] = \mathbb{P}[j \notin B_i(t)]$. Now let $A_{ji} = \inf\{t : j \in B_i(t)\}$. Note that $\{j \notin B_i(t)\} = \{A_{ji} > t\}$. Thus $\mathbb{P}[D_{ji} > t] = \mathbb{P}[A_{ji} > t]$, i.e. D_{ji} and A_{ji} are identically distributed. Now let $A_{(k)i} = \inf\{t : |B_i(t)| = k+1\}$. Note that $\{A_{(k)i} : 1 \leq k \leq n-1\}$ are the order statistics of $\{A_{ji} : j \in \{1, \dots, n\} \setminus \{i\}\}$.

If we reverse the arrow of time at time 0, then $B_i(t)$ is the set of infected users at time s in the SI epidemic spreading model where an infected user j contaminates a susceptible user ℓ at rate $by_{\ell j}$. It then follows that $W_{ki} \triangleq A_{(k)i} - A_{(k-1)i}$ is exponentially distributed with parameter μ_{ki} given by

$$\mu_{ki} = b \sum_{\substack{j \in B_i(A_{(k-1)i}) \\ \ell \notin B_i(A_{(k-1)i})}} y_{j\ell} \leq b|B_i(A_{(k-1)i})| = bk,$$

and hence $\mathbb{E}W_{ki} \geq (bk)^{-1}$. It follows that

$$\begin{aligned}
C_i &= \frac{1}{n-1} \sum_{j:j \neq i} \mathbb{E}D_{ji} = \frac{1}{n-1} \sum_{j:j \neq i} \mathbb{E}A_{ji} \\
&= \frac{1}{n-1} \sum_{k=1}^{n-1} \mathbb{E}A_{(k)i} = \frac{1}{n-1} \sum_{k=1}^{n-1} \sum_{\ell=1}^k \mathbb{E}W_{\ell i} \\
&\geq \frac{1}{n-1} \sum_{k=1}^{n-1} \sum_{\ell=1}^k \frac{1}{b\ell} = \frac{n}{b(n-1)} H_{n-1} - \frac{1}{b}.
\end{aligned}$$

The desired result follows by averaging over C_i . \square

4.3.3 Expander Network

We now consider a network characterized by an expander graph, which might be considered more realistic. An expander graph is a sparse graph with strong connectivity properties. An example is a d -regular graph which is often used in modeling social networks.

Finding a Nash equilibrium when the topology is an expander graph turns out to be quite complex. We thus consider *approximate* Nash equilibria. We define a user's strategy to be an ϵ -approximate NE if the cost to the user under this strategy is no worse than ϵ more than the cost of any other strategy [58]. More formally, the profile $\hat{\mathbf{y}}$ is an ϵ -approximate NE if for any k and any $\tilde{\mathbf{y}}_k \in \Delta_k$,

$$C_k(\hat{\mathbf{y}}) \leq \epsilon + C_k(\tilde{\mathbf{y}}_k, \hat{\mathbf{y}}_{-k}).$$

The ϵ -approximate price of stability is defined by (4.4) with Γ replaced by Γ_ϵ , the set of ϵ -approximate NE.

Suppose G is an expander network with bounded degree. We now show that the profile where users implement a uniform allocation of their budget of attention is an approximate NE.

Theorem 4.3.9. *In any network with homogeneous content creation rates λ and budgets of attention b , the uniform profile is a $\frac{d-1}{b}$ -approximate Nash equilibrium, where $d = \max_k d_k$ is the maximum degree of the graph.*

Proof. Consider user ℓ . For a given j ,

$$D_{j\ell}(\mathbf{y}_\ell, \mathbf{y}_{-\ell}) = \min_{k \in N(\ell)} \left\{ \frac{1}{y_{k\ell}} X_{k\ell} + Y_{jk}^\ell \right\},$$

where $\{X_{k\ell}\}$ are i.i.d. exponential random variable with parameter b , and Y_{jk}^ℓ is the time for an item originating from j to reach k without passing through ℓ . Define a random variable K by $K = \min\{k^* : Y_{jk^*}^\ell = \min_k Y_{jk}^\ell\}$. Since $\{Y_{jk}^\ell\}$ are independent of $\{X_{k\ell}\}$, so is K . Thus

$$\mathbb{E}X_{K\ell} = \sum_k \mathbb{E}[X_{k\ell}] \cdot \mathbb{P}[K = k] = b^{-1}.$$

We then have the following,

$$\mathbb{E}D_{j\ell}(\mathbf{u}_\ell, \mathbf{y}_{-\ell}) = \mathbb{E} \left[\min_{k \in N(\ell)} \{d_\ell X_{k\ell} + Y_{jk}^\ell\} \right] \leq \mathbb{E} [d_\ell X_{K\ell} + Y_{jk}^\ell] = \frac{d_\ell}{b} + \mathbb{E} \left[\min_{k \in N(\ell)} Y_{jk}^\ell \right].$$

On the other hand,

$$\begin{aligned} \mathbb{E}D_{j\ell}(\mathbf{y}_\ell, \mathbf{y}_{-\ell}) &= \mathbb{E} \left[\min_{k \in N(\ell)} \left\{ \frac{1}{y_{k\ell}} X_{k\ell} + Y_{jk}^\ell \right\} \right] \\ &\geq \mathbb{E} \left[\min_{k \in N(\ell)} \left\{ \frac{1}{y_{k\ell}} X_{k\ell} \right\} + \min_{k \in N(\ell)} Y_{jk}^\ell \right] = \frac{1}{b} + \mathbb{E} \left[\min_{k \in N(\ell)} Y_{jk}^\ell \right]. \end{aligned}$$

Thus

$$\mathbb{E}D_{j\ell}(\mathbf{u}_\ell, \mathbf{y}_{-\ell}) \leq \frac{d_\ell - 1}{b} + \mathbb{E}D_{j\ell}(\mathbf{y}_\ell, \mathbf{y}_{-\ell}).$$

Summing over j and then minimizing over \mathbf{y}_ℓ , we obtain

$$C_\ell(\mathbf{u}_\ell, \mathbf{y}_{-\ell}) \leq \frac{d_\ell - 1}{b} + \min_{\mathbf{y}_\ell \in \Delta_\ell} C_\ell(\mathbf{y}_\ell, \mathbf{y}_{-\ell}),$$

for any $\mathbf{y}_{-\ell}$. □

In view of Theorem 4.3.8, the next theorem shows that, for an expander network with edge expansion bounded away from zero, the uniform profile is order optimal, and hence the d -approximate price of stability is bounded.

Theorem 4.3.10. *The social cost C^u of the uniform profile is bounded by*

$$C^u \leq \frac{2d}{bh_G} H_{\lfloor n/2 \rfloor},$$

where $d = \max_i d_i$, and h_G is the edge expansion of G defined by

$$h_G = \min_{|A| \subset V} \frac{|\partial A|}{\min\{|A|, |A^c|\}},$$

with $\partial A = \{(u, v) \in E : u \in A, v \in A^c\}$.

Proof. The proof is essentially the same as that for Theorem 4.3.6. The difference is that the exponential random variable V_{jk} in (4.11) now has parameter μ_{jk} given by

$$\mu_{jk} = b \sum_{u \in F_{jk}, v \in F_{jk}^c} y_{uv} \geq \frac{b}{d} |\partial F_{jk}| \geq \frac{bh_G}{d} \min\{k, n - k\},$$

where F_{jk} is the set consisting of the first k users that has got the content originating from j , with $F_{j1} = \{j\}$. Hence

$$\sum_{i:i \neq j} \mathbb{E} D_{ji} \leq \sum_{k=1}^{n-1} \frac{(n-1)d}{bh_G \min\{k, n-k\}} \leq \frac{2d(n-1)}{bh_G} H_{\lfloor n/2 \rfloor}.$$

The desired result follows by summing over j and dividing by $n(n-1)$. □

4.4 Incentivizing Efficient Behavior

As we have just seen, topologies can be classified according to their performance under optimal and selfish allocations as

- (1). Efficient: These topologies have bounded PoS and optimal social cost of order $\log n$; they do not require additional mechanisms. Examples are expanders and cliques.
- (2). Inefficient amenable: These topologies have high PoS yet low (logarithmic) optimal social cost. As we shall show, incentive mechanisms can be introduced to change users' behavior and reduce their otherwise inefficient performance. Examples are the k -ary tree with bounded degree ($k = \Theta(1)$), and with low-scaling degree ($k = \Theta(n^\alpha)$, $\alpha \ll 1/2$).
- (3). Inefficient suboptimal: These topologies show inefficient content spread even under socially optimal allocations. No mechanism that preserves the topology and the budgets of attention can therefore lead to good performance. Examples are line and star networks, and k -ary trees with high-scaling degree ($k = \Theta(n^\alpha)$, α close to 1/2 and $\alpha > 1/2$).

We now propose an incentive mechanism that will prove particularly appealing for inefficient amenable graphs.

4.4.1 The Plus-One mechanism

Our mechanism relies on using incentives as a form of feedback for reallocating attention. Each receiver k , upon receiving useful information, sends a reward of 1, that we call a +1, to each node involved in relaying this information from its source s . Note that by useful information, we mean that the piece of information that arrived earliest at r , among all copies of this information at r . Therefore, a node sends a +1

to the neighbor through which the earliest copy was received. We now provide details of this mechanism.

- Each receiver k , upon reception of useful information from source s , sends a $+1_k$ to each node along the path to s that was involved in relaying that piece of information. Each node i along this path then keeps a score $O_j^i = \sum_{s,k:(j,i) \in \mathcal{P}_{s \leadsto k}^*} +1_k$, where $\mathcal{P}_{s \leadsto k}^*$ is the quickest (shortest) path from s to k . Note $\mathcal{P}_{s \leadsto k}^*$ is random and differs from one sample path to another.
 - The receiver is not required to know the topology of the network nor the path to each source. A completely distributed implementation consists of each receiver k sending a $+1_k$ to the neighbor through which it received the useful information from s . Each node along the path would then aggregate the $+1$ s it receives along with its own $+1$ before sending the sum up to its neighbor.
- At time intervals that are much longer than the slots over which $+1$ s are sent, each user i updates his allocation rates as follows,

$$y_{ji}(t+1) = y_{ji}(t) - \gamma_t \left(\delta_{ji}(t) - \frac{\sum_k \delta_{ki}(t)}{d_i} \right), \quad (4.12)$$

where

$$\delta_{ji}(t) = -\frac{1}{n(n-1)y_{ji}(t)^2} O_j^i,$$

and with γ_t such that $\sum_{t=0}^{\infty} \gamma_t = \infty$, $\lim_{t \rightarrow \infty} \gamma_t = 0$.

We refer to the $+1$ s as incentives because, as feedback, they represent the importance of a link, thus the value of the incentive to provide to bring about a favorable change in that link's allocation. In the present chapter we keep these incentives in quite general forms, but they may be regarded as monetary or non-monetary. Non-monetary incentives might include a form of reputation or recognition, such as in networks like Klout [2]. In such networks, users receive votes that count towards

their reputation or expertise, in return for some service (like answering questions) they provide to other users. A gain in reputation incites users to respond favorably when there is a possibility of receiving such votes. Note that such a method incentivizes a user to “serve” other users, thus going beyond a selfish allocation of attention.

The behavior induced by the Plus-One mechanism turns out to perform a stochastic gradient descent. The +1s collected by a node that correspond to a link e indeed serve to estimate the gradient of the cost with respect to the allocation on link e , x_e . We now show how δ_{ji} is an estimate of the gradient of the cost: $\frac{\partial C}{\partial y_{ji}}$. For ease of exposition we assume homogeneous budgets of attention ($b_i = 1$ for all i) and content creation rates ($\lambda_i = 1$ for all i). Let $\{X_e : e \in E\}$ be a collection of i.i.d. exponential random variable with parameter 1. The expected delay from source s to user k can be written as follows,

$$\mathbb{E}D_{sk}(\mathbf{y}) = \mathbb{E} \left[\sum_{e \in \mathcal{P}_{s \leadsto k}^*} \frac{1}{y_e} X_e \right].$$

Since the X_e 's are different with probability one, for each realization of the X_e 's, the shortest path $\mathcal{P}_{s \leadsto k}^*$ remains fixed for small enough perturbations in the y_e 's. The gradient of the average delay from s to k with respect to edge (j, i) can be estimated as follows,

$$\frac{\partial \mathbb{E}D_{sk}}{\partial y_{ji}} = \mathbb{E} \left[\mathbb{1}_{(j,i) \in \mathcal{P}_{s \leadsto k}^*} \frac{-1}{y_{ji}^2} X_{ji} \right],$$

where $\mathbb{1}_A$ is the indicator of A . From (4.3), the social cost is

$$C(\mathbf{y}) = \frac{1}{n(n-1)} \sum_{k \in V} \sum_{s: s \neq k} \mathbb{E}D_{sk}.$$

Thus the gradient of the overall cost is then given by

$$\frac{\partial C}{\partial y_{ji}} = -\frac{1}{n(n-1)} \mathbb{E} \left[\frac{X_{ji}}{y_{ji}^2} \sum_{s,k} \mathbb{1}_{(j,i) \in \mathcal{P}_{s \leadsto k}^*} \right].$$

An estimator $\tilde{\delta}_{ji}$ of $\frac{\partial C}{\partial y_{ji}}$ can then be written as follows,

$$\hat{\delta}_{ji} = -\frac{1}{n(n-1)y_{ji}^2} \sum_{s,k} \mathbb{1}_{(j,i) \in \mathcal{P}_{s \rightsquigarrow k}^*}.$$

Note that $\hat{\delta}_{ji}$ corresponds exactly to δ_{ji} , with $O_j^i = \sum_{s,k} \mathbb{1}_{(j,i) \in \mathcal{P}_{s \rightsquigarrow k}^*}$.

A study of the convergence properties of the Plus-One mechanism requires a careful analysis of the interchange of expectation and differentiation, which will not be pursued in the present chapter. Indeed results from simulations presented in Section 4.5 show convergence for all graphs that we consider.

4.4.2 Inefficient Suboptimal Graphs

The Plus-One mechanism performs well for the inefficient amenable graphs: the PoS is reduced and the cost under this distributed mechanism is very close to optimal, as we will see in Section 4.5. For inefficient suboptimal graphs, however, regardless of the PoS, the optimal social cost is still quite high. Our results from Section 4.3 show that line and star networks, and k -ary trees with $k = \Theta(n^\alpha)$, $\alpha > 0$, fall in this category of graphs. For such topologies, incentive mechanisms that promote only shifting of attention are not sufficient. More complex mechanisms that change the graph structure, or modify the budgets of attention of some nodes would seem more suitable. We leave the study of such mechanisms for future work.

4.5 Simulation Results

We now perform simulations to verify our efficiency results and validate the *Plus-One* (PO) mechanism. For each network topology, we ran a discrete-event simulation for three different scenarios. In the first scenario, the PO mechanism is implemented, so that the users are incentivized to minimize the social cost. In the second scenario, the users behave selfishly, so they optimize only for their own cost. The tuning of

their allocation of attention is similar to the PO mechanism except that since there is no reward from the downstream users, there is only a local optimization. In the first two scenarios, the initial allocation of attention is uniform for all users, which is reasonable without prior knowledge of the network. In the third scenario, the users do not tune their allocation, and maintain the uniform strategy.

We set the each user's publishing rate λ_i to $\lambda = 1$ and the budget of attention b_i to $b = 1$. The users update their allocation every 100 time units. We ran the simulation for length of time long enough so that the network reaches steady state as illustrated in Figure 4.1, which shows the average delay over time in a complete ternary tree with 1093 nodes. The average is taken within a window of 100 time units, the same as the interval between successive updates. We observe convergence around a small set of values in all scenarios for all topologies we consider.

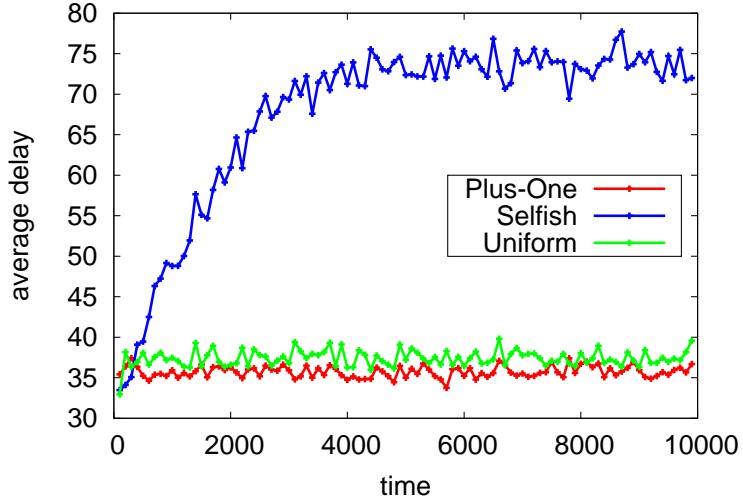


Figure 4.1: Average delay over time in a complete ternary tree with 1093 nodes.

We now study the steady-state average delay for various network topologies. For almost all cases we will plot the average delay derived from both theoretical and simulation results. Figure 4.2 plots the average delay over increasing network size for a line network. The PO mechanism indeed improves upon the selfish allocation,

achieving a cost close to the theoretical optimal social cost. Note however that the optimal social cost scales linearly with the network size. This line network is an example of inefficient suboptimal graphs that needs additional mechanisms beyond PO to improve the linear cost.

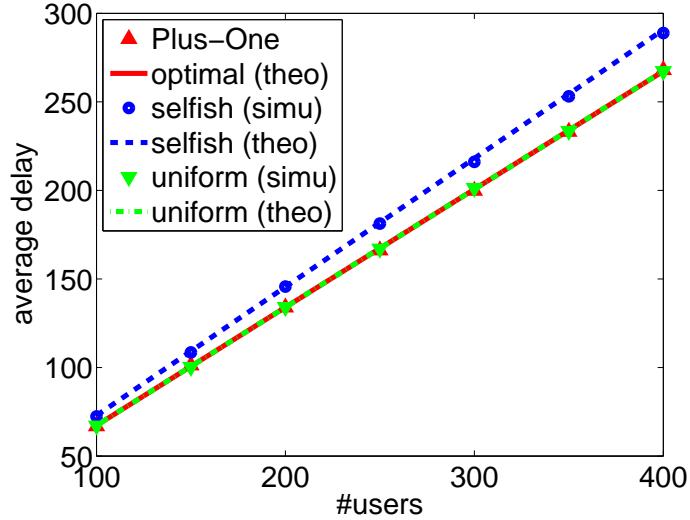


Figure 4.2: Average delay against network size for a line network.

Figure 4.3 plots the average delay against network size for complete k -ary trees with $k = 4$. The PO mechanism achieves the theoretical optimal social cost, which scales logarithmically with the network size. The social cost of the uniform strategy is only slightly higher than the optimal social cost, though the gap is increasing as the network size increases. In contrast, the cost for the selfish optimization is significantly higher, and increases much faster ($\sim n^{1/2}$) with the network size. This topology is an example of an inefficient amenable graph shows the power of the PO mechanism, in bringing the social cost down to one very close to the optimal cost.

Figure 4.4 plots the average delay in random 3-regular networks. Since random d -regular graphs have good expansion property with high probability, we know from Theorems 4.3.9 and 4.3.10 that the uniform strategy achieves a social cost that scales logarithmically, and that it is a 3-approximate NE. Figure 4.4 shows that the costs

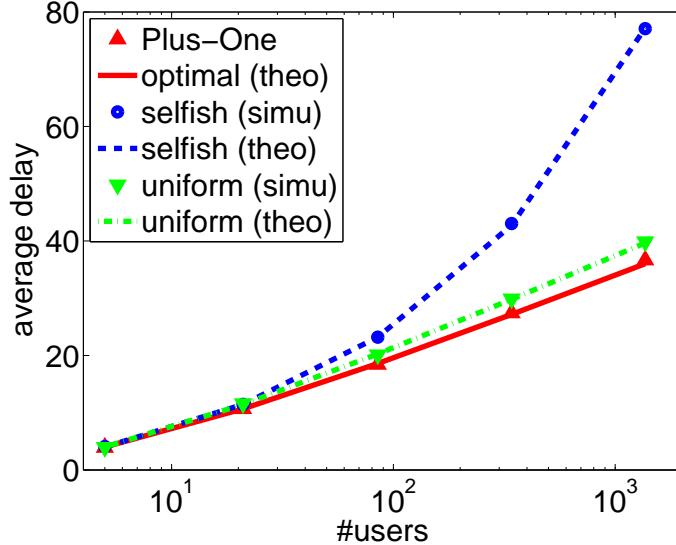


Figure 4.3: Average delay against network size for complete quaternary tree.

associated with PO, selfish optimization and the uniform strategy actually coincide for the network scales used in the simulation. This expander graph is an example of the efficient graphs, where the diversity of paths leads to optimal social costs without incentive mechanisms.

4.6 Conclusion and Future Work

This chapter has shown that social network topologies can be categorized into three classes according to efficiency of information spreading: efficient, inefficient amenable, and inefficient suboptimal. This chapter has also proposed the *Plus-One* mechanism, an incentive-based mechanism that brings the costs in inefficient amenable graphs close to optimal. Inefficient suboptimal graphs, on the other hand, are resilient to our mechanism, in that the cost under distributed attention allocation is reduced close to the optimal social cost, but the optimal social cost itself is quite high. For such graphs, mechanisms that go beyond incentives for attention shifting, those that induce a change in graph structure or in the budgets of attention are needed.

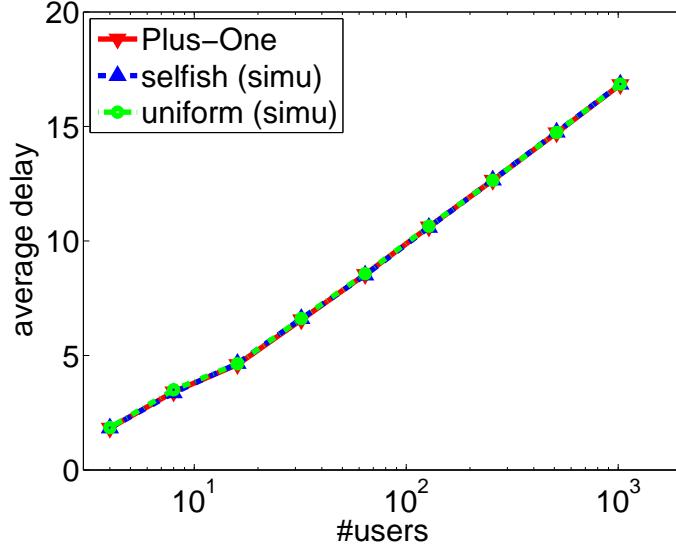


Figure 4.4: Average delay against network size for random 3-regular network.

We have demonstrated the effectiveness of the Plus-One mechanism by simulation. However, it will be of interest to have a formal investigation of its convergence property and provable performance guarantee.

We have assumed in the present chapter that all users are interested in receiving information from all sources. As an extension of the present chapter, we may consider a more interesting and realistic case where users have differing sets of interests. Mechanism design for such scenarios is decidedly more complex, with a more intricate contact structure.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

This thesis studied various aspects of networks characteristics and dynamics, with focus on reciprocity, competition and information dissemination.

Chapter 2 investigated the maximum reciprocity problem and its use with regard to the interpretation of empirical reciprocity in real networks. We proposed to interpret empirical reciprocity based on its comparison with the maximum possible reciprocity. We proved that the maximum reciprocity problem is NP-hard, so we did the comparison with an upper bound instead. We found that this bound is surprisingly close to the empirical reciprocity in a wide range of real networks, and that there is a surprisingly strong linear relationship between the two. We demonstrated that a particular type of small suboptimal motifs called 3-paths are the major cause for suboptimality in real networks.

There are several future directions related to Chapter 2. Given the usefulness and NP-hardness of the maximum reciprocity problem, it is of interest to design approximation algorithms with performance guarantee. The linear relationship between empirical reciprocity and the upper bound is intriguing and invites a careful study for its explanation. In Chapter 2, we fixed the degree sequence. It will be useful to consider small perturbation to the degree sequence and study the robustness of the results. We focused on maximum reciprocity with given degree constraints. To get a complete picture, it is necessary to study the full spectrum of reciprocity and in particular its minimum. More broadly, there are many other network characteristics of interest, whose interdependence is worthy of investigation.

Chapter 3 analyzes competition dynamics under cumulative advantage. We characterize the tail distributions of duration and intensity for pairwise competition under cumulative advantage. We demonstrate that duration always has a power-law tail irrespective of competitors’ fitness, while intensity has either a power-law tail or an exponential tail depending on whether the competitors are equally fit. We observe the struggle-of-the-fitness phenomenon, where a slight different in fitness results in an extremely heavy tail of duration distribution.

For future work, it is of interest to close the gap between the asymptotic upper and lower bounds for the duration distribution when the competitors are not equally fit. Another direction is to extend the results to more than two competitors and eventually to the full network setting.

Chapter 4 studied the efficiency of information dissemination in social networks with limited budget of attention. We quantified the efficiency of information dissemination for both cooperative and selfish user behaviors in various network topologies. We identified topologies where cooperation plays a critical role in efficient information propagation. We proposed an incentive mechanism called “plus-one” to coax users into cooperation in such cases, and demonstrated its effectiveness through simulation.

Future investigation is needed for the convergence property and provable performance guarantee of the “plus-one” mechanism. An analysis of power-law networks will be be valuable, since many social networks exhibit power-law degree distributions. Another natural extension is to allow users to have differing sets of interests.

APPENDIX A

ADDITIONAL PROOFS FOR CHAPTER 2

A.1 Proof of Theorem 2.3.7

We adapt the proof for Theorem 2.2 of [16] that deals with packing two graphic sequences for undirected graphs. Without loss of generality, we can assume that $V_0 = V$, since removing isolated vertices does not change the conclusion. Assume that conditions (1)–(3) hold and consider the set \mathcal{G} of all pairs of digraphs (G_1, G_2) such that

- (i). G_1 is symmetric with degree bi-sequence $(\mathbf{d}^0, \mathbf{d}^0)$,
- (ii). G_2 has degree bi-sequence $(\mathbf{d}^+ - \mathbf{d}^0, \mathbf{d}^- - \mathbf{d}^-)$,
- (iii). the union $G = G_1 + G_2$, as a multi-digraph, has degree bi-sequence $(\mathbf{d}^+, \mathbf{d}^-)$.

Note that G_1 can be identified with an undirected graph with degree sequence \mathbf{d}^0 . Conditions (1) and (2) guarantee that $\mathcal{G} \neq \emptyset$. Among all pairs in \mathcal{G} , choose a pair (G_1, G_2) such that the number of shared edges $|G_1 \cap G_2|$ is minimized. We will show that $G_1 \cap G_2 = \emptyset$, so their union $G = G_1 + G_2$ is a realization of $(\mathbf{d}^+, \mathbf{d}^-)$ and hence $\rho(\mathbf{d}^+, \mathbf{d}^-) \geq \rho(G) \geq |G_1| = 2 \sum_i d_i^0 \geq 2m$. To this end, we will show that condition (iii) would be violated if $G_1 \cap G_2 \neq \emptyset$.

Assume there exists an edge $(x, y) \in G_1 \cap G_2$. Since G_1 is required to be symmetric, for the sake of notational simplicity, we will use the same notation (a, b) for a single edge to refer to the pair of edges (a, b) and (b, a) in G_1 , which is represented pictorially by an undirected edge.

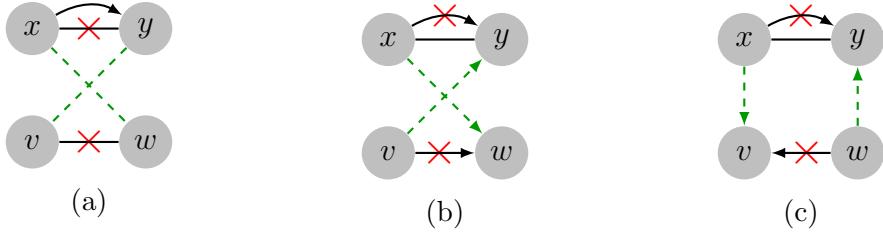


Figure A.1: Proof of Claims 1–3. The number of shared edges is reduced by rewiring the edges marked by red crosses into the dashed green edges. Each undirected edge represents a pair of reciprocated edges.

For $v \in V$, let $N_G(v) = \{u : (v, u) \in G\}$ be the out-neighbors of v in G and $N_G^-(v) = \{u : (u, v) \in G\}$ the in-neighbors of v in G . Let $N_G(v) = N_G^+(v) \cup N_G^-(v)$ be the neighbors of v in G . For $W \subset V$, let $N_G^+(W) = \bigcup_{w \in W} N_G^+(w)$, $N_G^-(W) = \bigcup_{w \in W} N_G^-(w)$ and $N_G(W) = N_G^+(W) \cup N_G^-(W)$. We use the convention $N_G^+(\emptyset) = N_G^-(\emptyset) = N_G(\emptyset) = \emptyset$. For $V_1, V_2 \subset V$, let $V_1 \otimes V_2 = \{(v_1, v_2) \in V_1 \times V_2 : v_1 \neq v_2\}$.

Now consider $I = V - [N_G(x) \cup N_G(y)]$. Let $W^1 = N_{G_1}(I)$, $W^2(I) = N_{G_2}^+(I) \cap N_{G_2}^-(I)$, $W^+ = N_{G_2}^+(I) - N_{G_2}^-(I)$ and $W^- = N_{G_2}^-(I) - N_{G_2}^+(I)$. Note that $N_G(I) = W^1 + W^2 + W^+ + W^-$.

We break the proof into several claims.

Claim 1. $W^1 \subset N_G(x) \cap N_G(y)$.

Proof. Suppose $W^1 \neq \emptyset$. Let $w \in W^1$ and $v \in I$ such that $(v, w) \in G_1$. If $w \notin N_G(x)$, then $G'_1 = G_1 - \{(x, y), (v, w)\} + \{(x, w), (y, v)\}$ would reduce the number of shared edges; see Figure A.1(a). Therefore, $w \in N_G(x)$. Similarly, $w \in N_G(y)$. \square

Claim 2. $N_{G_2}^+(I) \subset N_G^+(x)$.

Proof. Suppose $N_{G_2}^+(I) \neq \emptyset$. Let $w \in N_{G_2}^+(I)$ and $v \in I$ such that $(v, w) \in G_2$. If $w \notin N_G^+(x)$, then $G'_2 = G_2 - \{(x, y), (v, w)\} + \{(x, w), (v, y)\}$ would reduce the number of shared edges; see Figure A.1(b). Therefore, $w \in N_G^+(x)$. \square

Claim 3. $N_{G_2}^-(I) \subset N_G^-(y)$.

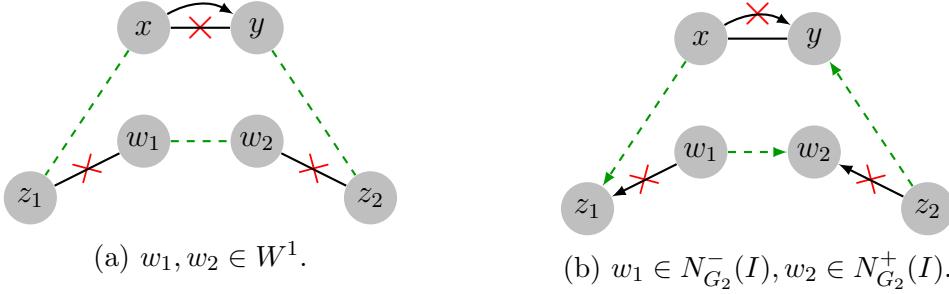


Figure A.2: Proof of Claims 4–5. The number of shared edges is reduced by rewiring the edges marked by red crosses into the dashed green edges. Each undirected edge represents a pair of reciprocal edges.

Proof. Suppose $N_{G_2}^-(I) \neq \emptyset$. Let $w \in N_{G_2}^-(I)$ and $v \in I$ such that $(w, v) \in G_2$. If $w \notin N_G^-(y)$, then $G'_2 = G_2 - \{(x, y), (v, w)\} + \{(x, v), (w, y)\}$ would reduce the number of shared edges; see Figure A.1(c). Therefore, $w \in N_G^-(y)$. \square

It follows from Claims 2 and 3 that $W^2 \subset N_G^+(x) \cap N_G^-(y)$, $W^+ \subset N_G^+(x)$ and $W^- \subset N_G^-(y)$. Note that $N_G(I) \subset N_G(x) \cup N_G(y) = I^c$. As a result, there is no edge with both ends in I .

Claim 4. *There exists an edge in G between every pair of distinct vertices in W^1 .*

Proof. Let $w_1, w_2 \in W^1$ and $z_1, z_2 \in I$ such that $(z_i, w_i) \in G_1$ for $i = 1, 2$, where z_1 and z_2 are not necessarily distinct. If $(w_1, w_2) \notin G$ and $(w_2, w_1) \notin G$, then $G'_1 = G_1 - \{(x, y), (w_1, z_1), (w_2, z_2)\} + \{(x, z_1), (y, z_2), (w_1, w_2)\}$ would reduce the number of shared edges; see Figure A.2(a). Therefore, either $(w_1, w_2) \in G$ or $(w_2, w_1) \in G$. \square

Claim 5. $N_{G_2}^-(I) \otimes N_{G_2}^+(I) \subset G$.

Proof. Let $(w_1, w_2) \in N_{G_2}^-(I) \otimes N_{G_2}^+(I)$. Let $z_1, z_2 \in I$ such that $(w_1, z_1) \in G_2$ and $(z_2, w_2) \in G_2$, where z_1 and z_2 are not necessarily distinct. If $(w_1, w_2) \notin G$, then $G'_2 = G_2 - \{(x, y), (w_1, z_1), (z_2, w_2)\} + \{(x, z_1), (z_2, y), (w_1, w_2)\}$ would reduce the number of shared edges; see Figure A.2(b). Therefore, $(w_1, w_2) \in G$. \square

As a result of Claim 5, $W^2 \otimes W^2 \subset G$ and $W^- \otimes W^+ \subset G$.

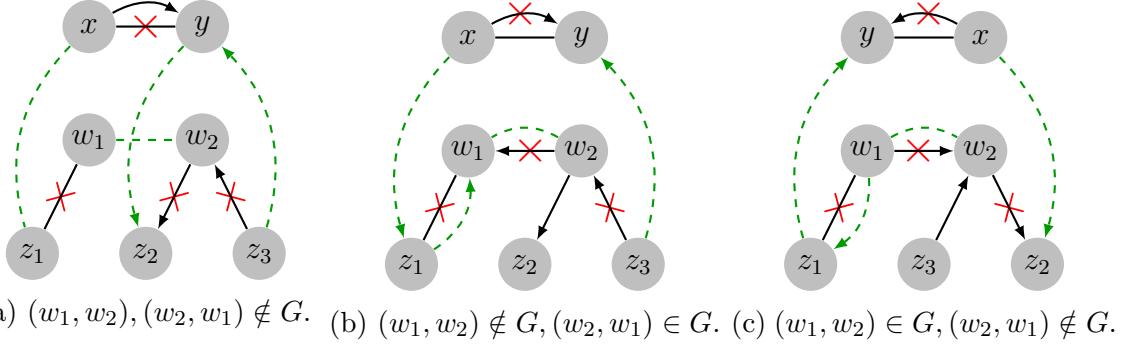


Figure A.3: Proof of Claim 6. The number of shared edges is reduced by rewiring the edges marked by red crosses into the dashed green edges, where $w_1 \in W^1$ and $w_2 \in W^2$. Each undirected edge represents a pair of reciprocal edges.

Claim 6. $W^1 \otimes W^2 \subset G$ and $W^2 \otimes W^1 \subset G$.

Proof. Let $w_1 \in W^1, w_2 \in W^2$ be such that $w_1 \neq w_2$. Let $z_1, z_2, z_3 \in I$ be such that $(w_1, z_1) \in G_1, (z_3, w_2) \in G_2$ and $(w_2, z_2) \in G_2$, where z_1, z_2, z_3 are not necessarily distinct. We will show that if $(w_1, w_2) \notin G$ or $(w_2, w_1) \notin G$, we would be able to find a new pair of graphs $(G'_1, G'_2) \in \mathcal{G}$ such that $|G'_1 \cap G'_2| < |G_1 \cap G_2|$, which would contradict the choice of (G_1, G_2) . Consider three cases.

(i). If $(w_1, w_2) \notin G, (w_2, w_1) \notin G$, then let

$$G'_1 = G_1 - \{(x, y), (w_1, z_1)\} + \{(x, z_1), (w_1, w_2)\},$$

$$G'_2 = G_2 - \{(z_3, w_2), (w_2, z_2)\} + \{(z_3, y), (y, z_2)\};$$

see Figure A.3(a).

(ii). If $(w_1, w_2) \notin G, (w_2, w_1) \in G$, then let

$$G'_1 = G_1 - \{(w_1, z_1)\} + \{(w_1, w_2)\},$$

$$G'_2 = G_2 - \{(x, y), (z_3, w_2), (w_2, w_1)\} + \{(z_3, y), (x, z_1), (z_1, w_1)\};$$

see Figure A.3(b).

(iii). If $(w_1, w_2) \in G$, $(w_2, w_1) \notin G$, then let

$$G'_1 = G_1 - \{(w_1, z_1)\} + \{(w_1, w_2)\},$$

$$G'_2 = G_2 - \{(x, y), (w_1, w_2), (w_2, z_2)\} + \{(x, z_2), (w_1, z_1), (z_1, x)\};$$

see Figure A.3(c).

Therefore, $(w_1, w_2) \in G$, $(w_2, w_1) \in G$. □

Claim 7. $\Delta \geq 3$.

Proof. Note that $\Delta \geq 2$ since $d_x^+ + d_x^- - d_x^0 \geq d_x^+ \geq 2$. If $\Delta = 2$, then $N_G(x) = \{y\}$ and $N_G(y) = \{x\}$. Thus $N_G(I) \subset \{x, y\}$. But $x, y \notin N_G(I)$ by the definition of I . Therefore, $N_G(I) = \emptyset$, and hence $d_v^+ = d_v^- = 0$ for every $v \in I = V - \{x, y\}$. It then follows that $\delta = \Delta = 2$ and $n = 2$. A direct calculation shows that condition (3) is violated. Therefore, $\Delta \geq 3$. □

Now consider the cut (I, I^c) . Let $W^0 = W^1 \cup W^2$. Let $w^0 = |W^0|$, $w^+ = |W^+|$, $w^- = |W^-|$. We will count the number of edges across the cut in a special way. An edge from G_2 in either direction is counted as one edge, while a pair of reciprocal edges from G_1 is counted as one edge. Note that the number of edges across the cut is bounded by

$$E(I, I^c) \geq \delta|I| = \delta(n - |I^c|),$$

since each vertex in I contributes at least δ edges. Note that $|N_G(x)| \leq \Delta - 1$ and $|N_G(y)| \leq \Delta - 1$, since the edge (x, y) has multiplicity 2. Since $W^0 \subset N_G(x) \cap N_G(y)$,

$$|I^c| = |N_G(x) \cup N_G(y)| = |N_G(x)| + |N_G(y)| - |N_G(x) \cap N_G(y)| \leq 2\Delta - 2 - w^0,$$

and hence

$$E(I, I^c) \geq \delta(n - 2\Delta + 2 + w^0).$$

On the other hand, each vertex in W^0 must connect to x, y and every other vertex in W^0 , and hence contributes at most $\Delta - 1 - w^0$ edges to $E(I, I^c)$. Each vertex in W^+ must connect to x and every vertex in W^- , and hence contributes at most $\Delta - 1 - w^-$ edges to $E(I, I^c)$. Similarly, every vertex in W^- contributes at most $\Delta - 1 - w^+$ edges to $E(I, I^c)$. Therefore,

$$E(I, I^c) \leq w^0(\Delta - 1 - w^0) + w^+(\Delta - 1 - w^-) + w^-(\Delta - 1 - w^+).$$

Combining the two inequalities for $E(I, I^c)$, we obtain

$$f(w^0, w^+, w^-) \geq \delta(n - 2\Delta + 2),$$

where

$$f(w^0, w^+, w^-) = w^0(\Delta - \delta - 1 - w^0) + w^+(\Delta - 1 - w^-) + w^-(\Delta - 1 - w^+).$$

Note that $w^0 + w^+ \leq |N_G(x) - \{y\}| \leq \Delta - 2$. Similarly $w^0 + w^- \leq \Delta - 2$. If we maximize f subject to these feasibility constraints, the inequality should still hold.

Claim 8. *The maximum value of $f(w^0, w^+, w^-)$ subject to the following constraints*

$$w^0 + w^+ \leq \Delta - 2,$$

$$w^0 + w^- \leq \Delta - 2,$$

$$w^0, w^+, w^- \geq 0,$$

is $f^* = (\Delta - 2)(\Delta - 1)$.

Proof. Note that f^* is achieved by $w^0 = w^- = 0$ and $w^+ = \Delta - 2$. Thus it remains to show that $f \leq f^*$ for all feasible (w^0, w^+, w^-) . For fixed w^0 and w^+ , f is linear in w^- , where $w^- \in [0, \Delta - 2 - w^0]$. Thus in order to maximize f , we only need to consider $w^- \in \{0, \Delta - 2 - w^0\}$. By the same argument, we only need to consider $w^+ \in \{0, \Delta - 2 - w^0\}$. Since f is symmetric in w^+ and w^- , we only need to consider three cases.

(i). $w^+ = w^- = 0$. In this case, $0 \leq w^0 \leq \Delta - 2$, and

$$f(w^0, 0, 0) = w^0(\Delta - \delta - 1 - w^0) \leq (\Delta - 2)(\Delta - \delta - 1) < f^*.$$

(ii). $w^- = 0$ and $w^+ = \Delta - 2 - w^0$. In this case,

$$f(w^0, w^+, 0) = w^0(\Delta - \delta - 1 - w^0) + (\Delta - 2 - w^0)(\Delta - 1) = f^* - w^0(w^0 + \delta) \leq f^*.$$

(iii). $w^+ = w^- = \Delta - 2 - w^0$. In this case,

$$\begin{aligned} f(w^0, w^+, w^-) &= w^0(\Delta - \delta - 1 - w^0) + 2(\Delta - 2 - w^0)(w^0 + 1) \\ &= w^0(3\Delta - \delta - 7 - 3w^0) + 2(\Delta - 2) \\ &\leq w^0(3\Delta - 8 - 3w^0) + 2(\Delta - 2). \end{aligned}$$

If $\Delta = 3$, then $w^0 \in \{0, 1\}$ and

$$f(w^0, w^+, w^-) \leq w^0(1 - 3w^0) + 2 \leq 2 = f^*.$$

If $\Delta \geq 4$, set $w^0 = (3\Delta - 8)/6$ and

$$f(w^0, w^+, w^-) \leq \frac{1}{12}(3\Delta - 8)^2 + 2(\Delta - 2).$$

Thus

$$f^* - f \geq (\Delta - 1)(\Delta - 2) - \frac{1}{12}(3\Delta - 8)^2 - 2(\Delta - 2) = \frac{1}{4}\Delta(\Delta - 4) + \frac{2}{3} \geq 0.$$

Therefore, $f \leq f^*$ for all feasible (w^0, w^+, w^-) , which completes the proof. \square

Now we have

$$(\Delta - 1)(\Delta - 2) = f^* \geq \delta(n - 2\Delta + 2),$$

and hence

$$\Delta \geq \sqrt{\delta n + \left(\delta - \frac{1}{2}\right)^2} + \frac{3}{2} - \delta.$$

which violates condition (3). Therefore, $G_1 \cap G_2 = \emptyset$ as desired.

A.2 Proof of Proposition 2.4.2

Let $G^{(i)}$ and $S^{(i)}$ be the digraph G and the set S before the i -th iteration of the **while** loop of lines 2–10. Given a vertex v , let $\Pi_v^{(i)}$ be the set of 3-paths in $G^{(i)}$ that starts at v , and $\tilde{\Pi}_v^{(i)} \subset \Pi_v^{(i)}$ the set of non-Type IV 3-paths. We first prove the following: If $\tilde{\Pi}_v^{(i)} \neq \emptyset$, then $v \in S^{(i)}$. This trivially holds for $i = 1$ since $S^{(1)} = V$. Assume it holds for the i -th iteration. Consider the $(i + 1)$ -st iteration. Let w_0 be such that $\tilde{\Pi}_{w_0}^{(i+1)} \neq \emptyset$. Let v_0 the node picked on line 3 of the i -th iteration. Consider two cases.

(1). Suppose the condition on line 4 is false. In this case, $G^{(i+1)} = G^{(i)}$ and hence

$$\tilde{\Pi}_{w_0}^{(i)} = \tilde{\Pi}_{w_0}^{(i+1)} \neq \emptyset. \text{ Thus } w_0 \neq v_0, \text{ and, by the induction hypothesis, } w_0 \in S^{(i)}.$$

By line 8, $S^{(i+1)} = S^{(i)} - \{v_0\}$, so $w_0 \in S^{(i+1)}$.

(2). Suppose the condition on line 4 is true. Let $\pi = (v_0, v_1, v_2, v_3)$ be the 3-path rewired in the i -th iteration. Since $S^{(i+1)} = S^{(i)} \cup \{v_1, v_2\}$ by line 6, by the

induction hypothesis, it suffices to show that $\Pi_{w_0}^{(i)} \neq \emptyset$ for $w_0 \notin \{v_0, v_1, v_2, v_3\}$.

Assume $w_0 \notin \{v_0, v_1, v_2, v_3\}$. If π is of Type III, then $G_a^{(i+1)} \subset G_a^{(i)}$ and hence $\emptyset \neq \tilde{\Pi}_{w_0}^{(i+1)} \subset \tilde{\Pi}_{w_0}^{(i)}$. Now suppose π is of Type I or II. Pick a 3-path $\pi_1 = (w_0, w_1, w_2, w_3) \in \tilde{\Pi}_{w_0}^{(i+1)}$. Note that the only edge in $G_a^{(i+1)} \setminus G_a^{(i)}$ is (v_0, v_3) . Since $(w_0, w_3) \notin G_a^{(i+1)}$ and $w_0 \neq v_0$, we obtain $(w_0, w_3) \notin G_a^{(i)}$. If $\pi_1 \in \Pi_{w_0}^{(i)}$, then $\pi_1 \in \tilde{\Pi}_{w_0}^{(i)}$. If $\pi_1 \notin \Pi_{w_0}^{(i)}$, then either (w_1, w_2) or (w_2, w_3) must be the newly added edge (v_0, v_3) . Suppose $(w_1, w_2) = (v_0, v_3)$. Then $\pi_2 = (w_0, w_1 = v_0, v_1, v_2) \in \Pi_{w_0}^{(i)}$. If $(w_0, v_2) \notin G_a^{(i)}$, then $\pi_2 \in \tilde{\Pi}_{w_0}^{(i)}$. If $(w_0, v_2) \in G_a^{(i)}$, then $\pi_3 = (w_0, v_2, v_3 = w_2, w_3) \in \Pi_{w_0}^{(i)}$. Since $(w_0, w_3) \notin G_a^{(i)}$, $\pi_3 \in \tilde{\Pi}_{w_0}^{(i)}$. Thus $\tilde{\Pi}_{w_0}^{(i)} \neq \emptyset$ if $(w_1, w_2) = (v_0, v_3)$. The same argument shows that $\tilde{\Pi}_{w_0}^{(i)} \neq \emptyset$ if $(w_2, w_3) = (v_0, v_3)$. Therefore, $\tilde{\Pi}_{w_0}^{(i)} \neq \emptyset$ for all cases.

Therefore, $\tilde{\Pi}_v^{(i)} \neq \emptyset$, then $v \in S^{(i)}$. When Algorithm 1 terminates, $S = \emptyset$, so there is no non-Type IV 3-paths. Now it remains to show that Algorithm 1 indeed terminates. For this purpose, let $X_i = |S^{(i+1)}| - |S^{(i)}|$. Let $Y_i = 1$ if the i -th iteration rewrites some 3-path and $Y_i = 0$ otherwise. Note that if $Y_i = 1$, $0 \leq X_i \leq 2$ and $|G_a^{(i+1)}| \leq |G_a^{(i)}| - 2$; otherwise, $X_i = -1$ and $|G_a^{(i+1)}| = |G_a^{(i)}|$. After the i -th iteration,

$$0 \leq |G_a^{(i+1)}| \leq |G_a^{(1)}| - 2 \sum_{j=1}^i Y_j,$$

and hence $2 \sum_{j=1}^i Y_j \leq |G_a^{(1)}| \leq |E|$. Thus

$$\begin{aligned} |S^{(i+1)}| &= |S^{(1)}| + \sum_{j=1}^i X_j \\ &\leq |V| + 2 \sum_{j=1}^i Y_j - \sum_{j=1}^i (1 - Y_i) \\ &= |V| - i + 3 \sum_{j=1}^i Y_j \\ &\leq |V| + \frac{3}{2}|E| - i. \end{aligned}$$

It follows that Algorithm 1 terminates in at most $|V| + \frac{3}{2}|E|$ iterations.

A.3 Proof of Lemma 2.4.8

We break the proof into several claims.

Claim 1. $(E_0 \cup E_1) \cap G_a^* = \emptyset$.

Proof. For $k < \ell$, let $\pi[v_k, v_\ell]$ be the sub-path of π from v_k to v_ℓ . Suppose there exists $(v_i, v_j) \in (E_0 \cup E_1) \cap G_a^*$. Note that $i \equiv j \pmod{2}$. If $i < j$, then $\pi[v_0, v_i] + (v_i, v_j) + \pi[v_j, v_{2p}]$ is a path of odd length $2p+1+i-j$, which requires $(v_0, v_{2p}) \in G_a^*$ by Lemma 2.4.3, a contradiction. If $i > j$, then $\pi[v_j, v_i] + (v_i, v_j)$ is a cycle in G_a^* . By Lemma 2.4.4, this must be a 3-cycle and $i = j+2$. By symmetry, we can assume $j \geq 1$. Lemma 2.4.5 applied to v_{j-1} and the 3-cycle (v_j, v_{j+1}, v_i, v_j) then requires $(v_{j-1}, v_{j+1}) \in G_a^*$, which we have just shown is impossible. Therefore, $(E_0 \cup E_1) \cap G_a^* = \emptyset$. \square

By virtue of Claim 1, a pair of edges (v_i, v_j) and (v_j, v_i) of $E_0 \cup E_1$ are either both in G^* or both outside G^* . Thus we only need to consider $(v_i, v_j) \in E_0 \cup E_1$ for $i < j$.

Claim 2. Either $(v_0, v_{2p}) \in G^*$ or $(v_1, v_{2p-1}) \in G^*$.

Proof. Suppose the contrary. By Claim 1, $(v_0, v_{2p}) \notin G_u^*$ and $(v_1, v_{2p-1}) \notin G_u^*$. Let $H = G^* - \{(v_0, v_1), (v_{2p-1}, v_{2p})\} + \{(v_0, v_{2p}), (v_{2p-1}, v_1)\}$. Then $\rho(H) = \rho(G^*)$ and hence H is also a maximum digraph. Now $\pi[v_1, v_{2p-1}] + (v_{2p-1}, v_1)$ is a $(2p-1)$ -cycle in H_a . If $p > 2$, this contradicts Lemma 2.4.4. If $p = 2$, this contradicts Lemma 2.4.5 since $(v_0, v_1) \notin H_a$ but $(v_0, v_{2p-1}) \in H_a$ by applying Lemma 2.4.3 to $\pi[v_0, v_{2p-1}]$. \square

Claim 3. If $(v_0, v_{2p}) \in G^*$, then $G^* \cap E_1 = \emptyset$ and $E_0 \subset G^*$.

Proof. Suppose $(v_{2i-1}, v_{2j-1}) \in G^*$, where $j > i \geq 1$. Note that $(v_{2i-1}, v_{2p}) \in G_a^*$ by Lemma 2.4.3. Then $C = \pi[v_0, v_{2j-1}] + (v_{2j-1}, v_{2i-1}, v_{2p}, v_0)$ satisfies the assumption of Lemma 2.4.6. Thus there exists an $H \in \mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$ with $\rho(H) = \rho(G^*) + 2(j-1) > \rho(G)$, a contradiction. Therefore, $(v_{2i-1}, v_{2j-1}) \notin G^*$ and hence $E_1 \cap G^* = \emptyset$.

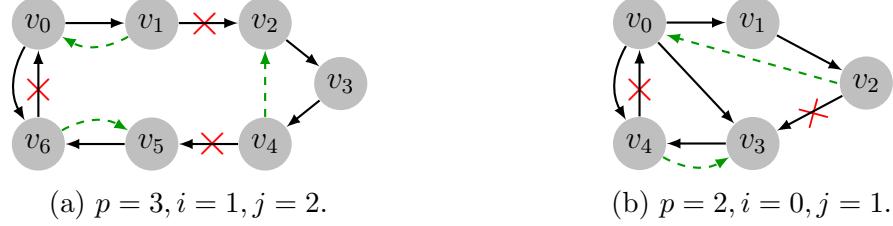


Figure A.4: Proof of Claim 3. Reciprocity can be increased by rewiring the edges marked by red crosses into the dashed green edges.

Suppose $(v_{2i}, v_{2j}) \notin G^*$, where $i < j$. Since $(v_0, v_{2p}) \in G^*$, either $i \geq 1$ or $j \leq p-1$.

By symmetry, we may assume that $j \leq p-1$. Let

$$H = G^* - \{(v_{2k-1}, v_{2k})\}_{k=1}^{i/2} - \{(v_{2k}, v_{2k+1})\}_{k=j/2}^{p-1} - \{(v_{2p}, v_0)\} \\ + \{(v_{2k-1}, v_{2k-2})\}_{k=1}^{i/2} + \{(v_{2k}, v_{2k-1})\}_{k=j/2+1}^p + \{(v_{2j}, v_{2i})\};$$

see Figures A.4. Then $H \in \mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$ and $\rho(H) = \rho(G^*) + 2i + 2(p-1-j) \geq \rho(G^*)$.

Thus H is also a maximum digraph, $i = 0$ and $j = p-1$. Now $(v_0, v_1, \dots, v_{2j}, v_0)$ is a cycle of length $2j+1 = 2p-1$ in H_a . This contradicts Lemma 2.4.4 if $p > 2$.

For $p = 2$, by applying Lemma 2.4.5 to v_3 and the 3-cycle (v_0, v_1, v_2, v_0) , we obtain $(v_1, v_3) \in H_a$, contradicting $E_1 \cap G^* = \emptyset$; see Figure A.4(b). Therefore, $(v_{2i}, v_{2j}) \in G^*$ and hence $E_0 \subset G^*$. \square

Claim 4. If $(v_1, v_{2p-1}) \in G^*$, then $G^* \cap E_0 = \emptyset$ and $E_1 \subset G^*$.

Proof. First consider the case $p \geq 3$. Claim 3 applied to $\pi[v_1, v_{2p-1}]$ yields $E_1 \subset G^*$ and $(v_{2i}, v_{2j}) \notin G^*$ for $i \geq 1$ and $j \leq p-1$. It remains to show $(v_{2i}, v_{2j}) \notin G^*$ for $i = 0$ or $j = p$. By symmetry, we only need to show $(v_0, v_{2j}) \notin G^*$. Suppose $(v_0, v_{2j}) \in G^*$. Consider the cycle $C = (v_0, v_3) + \pi[v_3, v_{2p-1}] + (v_{2p-1}, v_1, v_{2j}, v_0)$, which has length $2p$ and satisfies the assumption of Lemma 2.4.6. Note that $C \cap G_s^* = \{(v_{2p-1}, v_1), (v_{2j}, v_0)\}$. Lemma 2.4.6 then yields an $H \in \mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$ with $\rho(H) = \rho(G^*) + 2p - 4 > \rho(G^*)$, a contradiction. Thus $(v_0, v_{2j}) \notin G^*$ and $E_0 \cap G^* = \emptyset$.

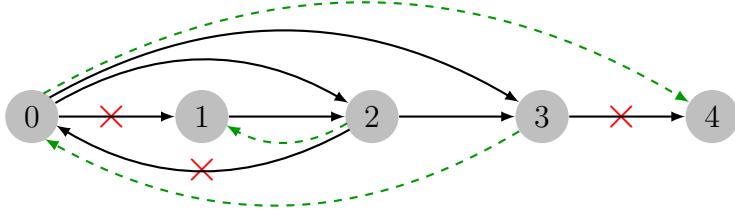


Figure A.5: Proof of Claim 4. Reciprocity can be increased by rewiring the edges marked by red crosses into the dashed green edges.

For $p = 2$, $E_1 \subset G^*$ trivially. By Claim 3, $(v_0, v_4) \notin G^*$. To show $E_0 \cap G^* = \emptyset$, by symmetry, we only need to show $(v_0, v_2) \notin G^*$. Suppose $(v_0, v_2) \in G^*$. Let

$$H = G^* - \{(v_0, v_1), (v_2, v_0), (v_3, v_4)\} + \{(v_3, v_0), (v_0, v_4), (v_2, v_1)\};$$

see Figure A.5. Then $H \in \mathcal{G}(\mathbf{d}^+, \mathbf{d}^-)$ and $\rho(H) = \rho(G) + 1$, a contradiction. Thus $(v_0, v_2) \notin G^*$ and $E_0 \cap G^* = \emptyset$. \square

APPENDIX B

DATASETS IN SECTION 2.5

Table B.1: Statistics of some real networks. The datasets without explicit citations are from the SNAP repository [52]. This table shows for each network the number of nodes (column 2), the number of edges (column 3), the number of reciprocal edges (column 4), the number of reciprocal edges in a 3-path optimal digraph returned by the GreedyRewire algorithm on page 20 (column 5), and the upper bound in Proposition 2.3.1 (column 6).

Network	Nodes	Edges	Reciprocal Edges		
			Observed	Algo. 1	Bound
Biological networks					
C. Elegan [73, 74]	297	2345	394	1364	1467
Mouse-Cortex [80]	49	964	656	804	825
Protein [70]	6339	34814	4216	22066	23630
Yeast [68]	6725	201775	1090	6446	8835
A. Thaliana [76]	10134	15580	12	40	77
Communication networks					
email-EuAll	265214	418956	108950	128192	143287
wiki-Talk	2394385	5021410	723690	1219196	1285201
Product co-purchasing networks					
amazon0302	262111	1234877	670170	801530	858907
amazon0312	400727	3200440	1701142	1945350	2079813
amazon0505	410236	3356824	1834774	2092700	2227333

Table B.1: (continued)

Network	Nodes	Edges	Reciprocal Edges		
			Observed	Algo. 1	Bound
amazon0601	403394	3387388	1887960	2130012	2266214
Social networks					
Epinions1	75879	508837	206194	299778	317821
Slashdot0811	77360	828161	717962	731044	737201
Slashdot0902	82168	870161	731862	749436	758751
Pokec	1632803	30622564	16641200	21997368	22813049
wiki-Vote	7115	103689	5854	31126	35989
LiveJournal1	4847571	68475391	51248308	55619590	56984610
LiveJournal [56]	5204176	76937805	56456064	61806458	63451685
Flickr [56]	1715255	22613980	14117878	16401174	16998181
YouTube [56]	1138499	4945382	3909878	3996410	4086949
Twitter [51]	41652230	1468364884	531703676	690897836	875520298
ego-Twitter	81306	1768135	851678	1112236	1179627
Google+ [35]	61858438	948605109	321728626	414578876	443168800
ego-Google+	107614	13673453	2870336	4954418	5481158
Stackoverflow [77]	1749197	11894846	26558	2445802	2965936
Web graphs					
BerkStan	685230	7600595	1902250	2257148	2913141
Google	875713	5105039	1565976	2106234	2460500
NotreDame	325729	1469679	759142	821340	907239
Stanford	281903	2312497	639722	770266	983414

Table B.1: (continued)

Network	Nodes	Edges	Reciprocal Edges		
			Observed	Algo. 1	Bound
Wikipedia [1]					
English	4709883	328267748	176523698	215049808	227103696
Swedish	1946669	49061638	10296750	12792974	13689733
Dutch	1794354	50061183	19993040	23471168	25078755
German	1738087	69385800	28079234	38594032	41799602
French	1555872	87231786	38347858	49859546	53102549
Russian	1163335	68613850	35807558	42472180	44437671
Italian	1160082	85261756	48584200	55921822	58593672
Spanish	1109589	32489175	4927794	10429430	11654906
Polish	1072883	51993365	28351902	32917546	34433059
Japanese	936882	61591797	26512542	36239836	38326442
Portuguese	841064	39840808	19062374	23016802	24224634
Chinese	781344	49703600	31848356	36082340	37248389
Korean	290291	15595628	9318976	10859386	11281173
P2P networks					
Gnutella04	10876	39994	0	13878	16371
Gnutella05	8846	31839	0	9584	11830
Gnutella06	8717	31525	0	9606	11825
Gnutella08	6301	20777	0	5604	6947
Gnutella09	8114	26013	0	7064	8822
Gnutella24	26518	65369	0	19142	23920
Gnutella25	22687	54705	0	15292	19016

Table B.1: (continued)

Network	Nodes	Edges	Reciprocal Edges		
			Observed	Algo. 1	Bound
Gnutella30	36682	88328	0	25386	31236
Gnutella31	62586	147892	0	40564	50227
Call Graph [66]					
DrJava	1702	2920	4	778	1056
Endeavour	724	2067	2	358	519
FreeMind	237	623	18	140	217
JabRef	868	1532	2	340	523
jEdit	2222	5172	10	1286	1793
JForum	716	1506	2	248	364
JPetStore	222	328	0	30	42
Kunagi	781	1345	6	348	599
logicaldoc	892	3682	0	194	304
Makagiga	1777	4075	8	1106	1440
OpenKM	1390	2525	0	384	491
openproj	2824	4866	2	1428	1823
OpenSyncro	658	1271	2	216	327
SweetHome3D	1118	2363	12	558	844
weka	911	1737	2	392	581
Linux [75]	12391	33553	316	7982	10933

APPENDIX C

ADDITIONAL PROOFS FOR CHAPTER 3

This appendix provides proofs for the main results in Section 3.4. We will use the following additional notations and definitions.

- Denote the transition probability from state (x_0, y_0) to state (x, y) in $t = x + y - x_0 - y_0$ steps by

$$p_r(x_0, y_0; x, y) = \mathbb{P}_{\text{CA},r}^{(x_0, y_0)}[(X_t, Y_t) = (x, y)].$$

Note that the transition probability is nonzero only for this specific t . Thus we will often omit to mention t explicitly hereafter and assume that the appropriate t has been chosen.

- Let τ_n be the time of the n -th tie, which can be defined recursively by $\tau_0 = -\infty$ and

$$\tau_n = \inf\{t > \tau_{n-1} : X_t = Y_t\}, \quad n \geq 1.$$

Note that $T = \tau_N$.

- Denote by $q_r(x_0, y_0)$ the probability of having no tie after leaving state (x_0, y_0) , i.e.,

$$q_r(x_0, y_0) = \mathbb{P}_{\text{CA},r}^{(x_0, y_0)}[X_t \neq Y_t, t \geq 1].$$

Note that $q_r(x_0, x_0) = \mathbb{P}_{\text{CA},r}^{(x_0, x_0)}[\tau_2 = \infty]$ and $q_r(x_0, y_0) = \mathbb{P}_{\text{CA},r}^{(x_0, y_0)}[\tau_1 = \infty]$ for $x_0 \neq y_0$.

- Denote by $A_{n,t}(x,y)$ the set of paths that start from (x,y) at time 0 and end with the n -th tie at time t , i.e., $\tau_n = t$.

C.1 Proof of Theorem 3.4.1

Note that the CA₌ model is the standard Pólya urn model. The proof of Theorem 3.4.1 combines known results for this model. Starting from the initial state (x_0, y_0) , X_t has a beta-binomial distribution with parameters x_0 and y_0 [45]. Note that the event $X_t = Y_t$ occurs only if $t = |x_0 - y_0| + 2k$ for some integer $k \geq 0$. For such t , $X_t = Y_t$ if and only if $X_t = z_0 + k$, where $z_0 = \max\{x_0, y_0\}$. By Eq. (6.27) of [45],

$$\mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[X_t = Y_t] = \mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[X_t = z_0 + k] = \frac{B(z_0 + k, z_0 + k)}{B(x_0, y_0)} \binom{t}{k}. \quad (\text{C.1})$$

Recall that $q_1(x, y)$ is the probability of having no tie after leaving state (x, y) . Thus

$$\mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[T = t] = \mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[X_t = Y_t] \cdot q_1(z_0 + k, z_0 + k), \quad (\text{C.2})$$

where the second factor on the right-hand side is the probability of having no tie after t .

Recall that the exit probability $E(x, y)$ in [6] is the probability of ever having a tie starting from (x, y) , including the initial state (x, y) . Thus for $x \neq y$, $q_1(x, y)$ is related to $E(x, y)$ by

$$q_1(x, y) = 1 - E(x, y).$$

Using Eq. (22) of [6] for $E(x, y)$, we obtain

$$q_1(x+1, x) = q_1(x, x+1) = \frac{\Gamma(x+1/2)}{\Gamma(x+1)\Gamma(1/2)}.$$

However, $q_1(x, x) \neq E(x, x) = 1$. By considering the one-step transition from (x, x) to $(x + 1, x)$ or $(x, x + 1)$, we obtain

$$q_1(x, x) = \frac{1}{2}q_1(x + 1, x) + \frac{1}{2}q_1(x, x + 1) = \frac{\Gamma(x + 1/2)}{\Gamma(x + 1)\Gamma(1/2)}.$$

Eliminating $\Gamma(x + 1/2)$ by the identity

$$\Gamma(2x) = \pi^{-1/2}2^{2x-1}\Gamma(x)\Gamma(x + 1/2)$$

in [60, Eq. (5.5.5)], and using $\Gamma(x + 1) = x\Gamma(x)$ and $\Gamma(1/2) = \sqrt{\pi}$, we obtain

$$q_1(x, x) = \frac{\Gamma(2x)}{x2^{2x-1}\Gamma(x)\Gamma(x)} = \frac{1}{x2^{2x-1}B(x, x)}. \quad (\text{C.3})$$

Substitution of Eqs. (C.1) and (C.3) into Eq. (C.2) yields

$$\mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[T = t] = \frac{1}{B(x_0, y_0)} \cdot \frac{1}{(z_0 + k)2^{2k+2z_0-1}} \binom{t}{k}.$$

For $t = |x_0 - y_0| + 2k$, Stirling's formula yields

$$\binom{t}{k} \sim \sqrt{\frac{2}{\pi}} t^{-1/2} 2^t,$$

and hence

$$\mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[T = t] \sim \frac{1}{2^{x_0+y_0-5/2}\sqrt{\pi}B(x_0, y_0)} t^{-3/2}. \quad (\text{C.4})$$

It is well-known that as $t \rightarrow \infty$, $X_t/(X_t + Y_t)$ converges almost surely to a beta random variable V . It follows that $|X_t - Y_t|/(X_t + Y_t) \rightarrow |2V - 1|$. Thus, for $V \neq 1/2$, which holds almost surely, we have $|X_t - Y_t|/(X_t + Y_t) > 0$ for all large enough t . Therefore, $\mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[T = \infty] = 0$. Summing over t in Eq. (C.4), we obtain as $t \rightarrow \infty$,

$$\mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[T \geq t] = \sum_{t'=t}^{\infty} \mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[T = t']$$

$$\begin{aligned}
&\sim \frac{1}{2} \sum_{s=t}^{\infty} \frac{1}{2^{x_0+y_0-5/2} \sqrt{\pi} B(x_0, y_0)} s^{-3/2} \\
&\sim \frac{1}{2} \int_t^{\infty} \frac{1}{2^{x_0+y_0-5/2} \sqrt{\pi} B(x_0, y_0)} s^{-3/2} ds \\
&= \frac{1}{2^{x_0+y_0-5/2} \sqrt{\pi} B(x_0, y_0)} t^{-1/2},
\end{aligned}$$

where we have used the fact that half of the terms are zero in the second step, and $\sum_{s=t}^{\infty} s^{-a} \sim \int_t^{\infty} s^{-a} ds$ in the third step. This completes the proof of Theorem 3.4.1.

C.2 Proof of Theorem 3.4.2

Similar to Eq. (C.2), we have

$$\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[T = t] = p_r(x_0, y_0; z_0 + k, z_0 + k) \cdot q_r(z_0 + k, z_0 + k). \quad (\text{C.5})$$

Thus the proof here amounts to finding expressions for both $p_r(x_0, y_0; z_0 + k, z_0 + k)$ and $q_r(z_0 + k, z_0 + k)$ in CA_{\neq} . We break the proof into three lemmas.

Lemma C.2.1.

$$p_r(x_0, y_0; x_0 + k, y_0 + h) \geq \frac{(x_0)_k (y_0)_h}{(rx_0 + y_0)_{k+h}} \binom{k+h}{k}, \quad (\text{C.6})$$

for all $k \geq 0, h \geq 0$.

Lemma C.2.2.

$$p_r(x_0, y_0; x_0 + k, y_0 + h) \leq \frac{(x_0)_k (y_0)_h}{(r^{-1})_k (rx_0 + y_0)_h}, \quad (\text{C.7})$$

for all $k \geq 0, h \geq 0$.

Lemma C.2.3. For $r > 1$,

$$q_r(x, x) \rightarrow \frac{r-1}{r+1}, \quad (\text{C.8})$$

as $x \rightarrow \infty$.

Before proving these lemmas, we first use them to prove Theorem 3.4.2.

Proof of Theorem 3.4.2. By Lemma C.2.1, we have

$$\begin{aligned}
& p_r(x_0, y_0; z_0 + k, z_0 + k) \\
& \geq \frac{(x_0)_{k+z_0-x_0} (y_0)_{k+z_0-y_0}}{(rx_0 + y_0)_{2k+2z_0-x_0-y_0}} \binom{2k + 2z_0 - x_0 - y_0}{k + z_0 - x_0} \\
& = \frac{\Gamma(rx_0 + y_0)}{\Gamma(x_0)\Gamma(y_0)} \cdot \frac{\Gamma(k+z_0)\Gamma(k+z_0)}{\Gamma(k+z_0-x_0+1)\Gamma(k+z_0-y_0+1)} \cdot \frac{\Gamma(2k+2z_0-x_0-y_0+1)}{\Gamma(2k+2z_0+(r-1)x_0)}
\end{aligned}$$

Using the relation $\Gamma(k+a)/\Gamma(k+b) \sim k^{a-b}$ as $k \rightarrow \infty$, we obtain

$$\begin{aligned}
p_r(x_0, y_0; z_0 + k, z_0 + k) & \gtrsim \frac{\Gamma(rx_0 + y_0)}{2^{x_0+y_0-2}\Gamma(x_0)\Gamma(y_0)} (2k)^{-(r-1)x_0-1} \\
& = 2(r+1)x_0\varphi_1(2k)^{-(r-1)x_0-1},
\end{aligned}$$

where φ_1 is given by Eq. (3.4). Application of this asymptotic bound and Lemma C.2.3 to Eq. (C.5) yields

$$\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[T = t] \gtrsim 2(r-1)x_0\varphi_1 t^{-(r-1)x_0-1}. \quad (\text{C.9})$$

Note that $\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[T = +\infty] = \mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[N = +\infty] = 0$, where the second equality will follow from Theorem 3.4.4, so we will not provide a separate proof here. Summing over t in Eq. (C.9) and noting that half of the terms are zero, we obtain as $t \rightarrow \infty$,

$$\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[T \geq t] = \sum_{t'=t}^{\infty} \mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[T = t'] \gtrsim \int_t^{\infty} (r-1)x_0\varphi_1 s^{-(r-1)x_0-1} ds = \varphi_1 t^{-(r-1)x_0},$$

establishing the lower bound.

In a similar way, Lemma C.2.2 yields

$$p_r(x_0, y_0; z_0 + k, z_0 + k) \lesssim 2(r+1)(x_0 - r^{-1})\varphi_2(2k)^{-(r-1)(x_0 - r^{-1})-1},$$

where φ_2 is given by Eq. (3.5). Application of this asymptotic bound and Lemma C.2.3 to Eq. (C.5) yields

$$\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[T = t] \lesssim 2(r-1)(x_0 - r^{-1})\varphi_2 t^{-(r-1)(x_0 - r^{-1})-1},$$

and

$$\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[T \geq t] = \sum_{t'=t}^{\infty} \mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[T = t'] \lesssim \varphi_2 t^{-(r-1)(x_0 - r^{-1})},$$

establishing the upper bound. \square

Now we prove the lemmas. Recall that the transition probability $p(x_0, y_0; x, y)$ of going from (x_0, y_0) to (x, y) satisfies the following recursion (Chapman-Kolmogorov equation),

$$p_r(x_0, y_0; x, y) = \frac{r(x-1)}{r(x-1) + y} p_r(x_0, y_0; x-1, y) + \frac{y-1}{rx + y - 1} p_r(x_0, y_0; x, y-1), \quad (\text{C.10})$$

for $x \geq x_0$, $y \geq y_0$ and $x+y \geq x_0+y_0+1$, with the boundary condition $p_r(x_0, y_0; x, y) = 0$ for $x < x_0$ or $y < y_0$. Note that we have replaced the one-step transition probabilities $Q_{\text{CA},r}(x-1, y; x, y)$ and $Q_{\text{CA},r}(x, y-1; x, y)$ by the expressions in Eq. (3.1).

Proof of Lemma C.2.1. We will use the short-hand notation $p(k, h)$ for $p_r(x_0, y_0; x_0 + k, y_0 + h)$, and $\psi(k, h)$ for the right-hand side of Eq. (C.6). We first prove the boundary case for $k = 0$. By Eq. (C.10), for $h \geq 1$,

$$p(0, h) = \frac{y_0 + h - 1}{rx_0 + y_0 + h - 1} p(0, h-1),$$

which is a simple recursion in h and can be expanded to yield

$$p(0, h) = \frac{(y_0)_h}{(rx_0 + y_0)_h} p(0, 0) = \frac{(y_0)_h}{(rx_0 + y_0)_h} = \psi(0, h),$$

which yields Eq. (C.6) for $k = 0$ and $h \geq 1$. Here we have used $p(0, 0) = p_r(x_0, y_0; x_0, y_0) = 1$.

Similarly, for the other boundary case $h = 0, k \geq 1$, we have

$$p(k, 0) = \frac{(x_0)_k}{(x_0 + r^{-1}y_0)_k} p(0, 0) = \frac{(x_0)_k}{(x_0 + r^{-1}y_0)_k} \geq \frac{(x_0)_k}{(rx_0 + y_0)_k} = \psi(k, 0),$$

where the last inequality is because $(x)_k$ increases with x , and $x_0 + r^{-1}y_0 \leq rx_0 + y_0$.

For the general case, we use induction on $k+h$. The base case $k+h = 1$ is already proven, since either $k = 0$ or $h = 0$ when $k+h = 1$. Assume Eq. (C.6) holds for $k+h = m \geq 1$. Consider $k+h = m+1$. We can also assume $k \geq 1$ and $h \geq 1$, since we have proven the boundary cases for $k = 0$ or $h = 0$. The recursion in Eq. (C.10) yields

$$\begin{aligned} p(k, h) &= \frac{r(x_0 + k - 1)}{rk + h + c_0 - r} p(k-1, h) + \frac{y_0 + h - 1}{rk + h + c_0 - 1} p(k, h-1) \\ &\geq \frac{r(x_0 + k - 1)}{rk + h + c_0 - 1} p(k-1, h) + \frac{y_0 + h - 1}{rk + h + c_0 - 1} p(k, h-1), \end{aligned}$$

where $c_0 = rx_0 + y_0$.

Applying the induction hypothesis $p(k-1, h) \geq \psi(k-1, h)$ and $p(k, h-1) \geq \psi(k, h-1)$ to the above inequality yields

$$\begin{aligned} p(k, h) &\geq \frac{r(x_0 + k - 1)}{rk + h + c_0 - 1} \psi(k-1, h) + \frac{y_0 + h - 1}{rk + h + c_0 - 1} \psi(k, h-1) \\ &= \frac{(rk + h)(k + h + c_0 - 1)}{(k + h)(rk + h + c_0 - 1)} \psi(k, h), \end{aligned}$$

where in the last step we have used

$$\psi(k-1, h) = \frac{k}{k+h} \cdot \frac{k+h+c_0-1}{x_0+k-1} \psi(k, h),$$

and

$$\psi(k, h-1) = \frac{h}{k+h} \cdot \frac{k+h+c_0-1}{y_0+h-1} \psi(k, h).$$

To complete the proof, it suffices to show that

$$\frac{(rk+h)(k+h+c_0-1)}{(k+h)(rk+h+c_0-1)} \geq 1,$$

but this is equivalent to $r \geq 1$, which is true by assumption. \square

Proof of Lemma C.2.2. The proof of Lemma C.2.2 follows the same line of reasoning as that used to prove Lemma C.2.1. The boundary cases can be verified directly. We only outline the induction step here. Applying Eq. (C.7) to the right-hand side of Eq. (C.10) yields

$$\begin{aligned} p(k, h) &\leq \frac{r(x_0+k-1)}{rk+h+c_0-r} \cdot \frac{(x_0)_{k-1}(y_0)_h}{(r^{-1})_{k-1}(c_0)_h} + \frac{y_0+h-1}{rk+h+c_0-1} \cdot \frac{(x_0)_k(y_0)_{h-1}}{(r^{-1})_k(c_0)_{h-1}} \\ &= \left[\frac{r(r^{-1}+k-1)}{rk+h+c_0-r} + \frac{c_0+h-1}{rk+h+c_0-1} \right] \frac{(x_0)_k(y_0)_h}{(r^{-1})_k(c_0)_h}. \end{aligned}$$

Note that

$$\frac{r(r^{-1}+k-1)}{rk+h+c_0-r} + \frac{c_0+h-1}{rk+h+c_0-1} \leq \frac{r(r^{-1}+k-1)}{rk+h+c_0-r} + \frac{c_0+h-1}{rk+h+c_0-r} = 1,$$

which completes the induction. \square

Proof of Lemma C.2.3. Recall that $A_{n,2k}(x, x)$ is the set of paths that start from (x, x) at time 0 and end with the n -th tie at time $2k$, i.e., $\tau_n = 2k$. Let $A_{n,2k} = A_{n,2k}(0, 0)$.

Note that the paths in $A_{n,2k}(x, x)$ are exactly the paths in $A_{n,2k}$ translated by (x, x) . Let $\tilde{\pi} \in A_{n,2k}$ and its state at time t be $\tilde{\pi}_t = (\tilde{x}_t, \tilde{y}_t)$. The translation of $\tilde{\pi}$ by (x, x) , denoted $x + \tilde{\pi}$, is a path in $A_{n,2k}(x, x)$, whose probability in the CA model is given by

$$\mathbb{P}_{\text{CA},r}^{(x,x)}[x + \tilde{\pi}] = \prod_{j=0}^{2k-1} \left(\frac{r(x + \tilde{x}_j)}{r(x + \tilde{x}_j) + (x + \tilde{y}_j)} \right)^{\tilde{x}_{j+1} - \tilde{x}_j} \left(\frac{(x + \tilde{y}_j)}{r(x + \tilde{x}_j) + (x + \tilde{y}_j)} \right)^{\tilde{y}_{j+1} - \tilde{y}_j}.$$

For fixed k and $\tilde{\pi}$, as $x \rightarrow \infty$, $\mathbb{P}_{\text{CA},r}^{(x,x)}[x + \tilde{\pi}]$ converges to

$$\prod_{j=0}^{2k-1} \left(\frac{r}{r+1} \right)^{\tilde{x}_{j+1} - \tilde{x}_j} \left(\frac{1}{r+1} \right)^{\tilde{y}_{j+1} - \tilde{y}_j} = \mathbb{P}_{\text{RW},r}^{(0,0)}[\tilde{\pi}],$$

which corresponds to the probability of the path $\tilde{\pi}$ in a random walk with parameter $r/(r+1)$. Thus, as $x \rightarrow \infty$,

$$\mathbb{P}_{\text{CA},r}^{(x,x)}[\tau_n = 2k] = \sum_{\tilde{\pi} \in A_{n,2k}} \mathbb{P}_{\text{CA},r}^{(x,x)}[x + \tilde{\pi}] \rightarrow \sum_{\tilde{\pi} \in A_{n,2k}} \mathbb{P}_{\text{RW},r}^{(0,0)}[\tilde{\pi}] = \mathbb{P}_{\text{RW},r}^{(0,0)}[\tau_n = 2k].$$

After summing over k and using the Dominated Convergence Theorem, we obtain

$$\mathbb{P}_{\text{CA},r}^{(x,x)}[\tau_n < \infty] = \sum_{k=1}^{\infty} \mathbb{P}_{\text{CA},r}^{(x,x)}[\tau_n = 2k] \rightarrow \sum_{k=1}^{\infty} \mathbb{P}_{\text{RW},r}^{(0,0)}[\tau_n = 2k] = \mathbb{P}_{\text{RW},r}^{(0,0)}[\tau_n < \infty].$$

In particular,

$$q_r(x, x) = 1 - \mathbb{P}_{\text{CA},r}^{(x,x)}[\tau_2 < \infty] \rightarrow 1 - \mathbb{P}_{\text{RW},r}^{(0,0)}[\tau_2 < \infty] = \frac{r-1}{r+1},$$

where we have used Eq. (3.3) in [38] for $\mathbb{P}_{\text{RW},r}^{(0,0)}[\tau_2 < \infty]$ in the last step. \square

C.3 Proof of Theorem 3.4.3

Recall that $A_{n,t}(x_0, y_0)$ is the set of paths starting from (x_0, y_0) that end with the n -th tie at time t , i.e., $\tau_n = t$. We will use the short-hand notation $A_{n,t}$ for $A_{n,t}(x_0, y_0)$. As in Section C.1, the set $A_{n,t}$ is non-empty only if $t = |x_0 - y_0| + 2k$ for some integer $k \geq n - 1$, in which case, every path in $A_{n,t}$ ends in state $(z_0 + k, z_0 + k)$ with $z_0 = \max\{x_0, y_0\}$. Recall from [6] that the probability of any path π connecting states (x_0, y_0) and (x, y) is

$$\mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[\pi] = \frac{B(x, y)}{B(x_0, y_0)} = \frac{B(x, y)}{B(x_0, y_0)} 2^t \mathbb{P}_{\text{RW},1}^{(x_0,y_0)}[\pi].$$

Summing over $\pi \in A_{n,t}$, where $x = y = z_0 + k$, we obtain

$$\mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[\tau_n = t] = \frac{B(z_0 + k, z_0 + k)}{B(x_0, y_0)} 2^t \mathbb{P}_{\text{RW},1}^{(x_0,y_0)}[\tau_n = t]. \quad (\text{C.11})$$

Thus the probability of having the n -th and also the last tie at time t is given by

$$\begin{aligned} \mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[T = t, N = n] &= \mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[\tau_n = t] \cdot q_1(z_0 + k, z_0 + k) \\ &= \frac{1}{2^{x_0+y_0-2} B(x_0, y_0)} \cdot \frac{1}{t + x_0 + y_0} \mathbb{P}_{\text{RW},1}^{(x_0,y_0)}[\tau_n = t], \end{aligned} \quad (\text{C.12})$$

where we have used Eqs. (C.11) and (C.3) in the last step. Note that $\mathbb{P}_{\text{RW},1}^{(x_0,y_0)}[\tau_n = t]$ is the probability $f_{n,t}(d_0)$ of the n -th visit to the origin at time t in a simple symmetric random walk starting from $d_0 = |x_0 - y_0|$. Summing over t in (C.12), we obtain

$$\begin{aligned} \mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[N = n] &= \frac{1}{2^{x_0+y_0-2} B(x_0, y_0)} \sum_{k=n-1}^{\infty} \frac{1}{2k + d_0 + x_0 + y_0} f_{n,d_0+2k}(d_0) \\ &= \frac{1}{2^{x_0+y_0-2} B(x_0, y_0)} G_n(1; d_0), \end{aligned} \quad (\text{C.13})$$

where

$$G_n(z; d_0) = \sum_{k=n-1}^{\infty} \frac{1}{2k + d_0 + x_0 + y_0} f_{n,d_0+2k}(d_0) z^{d_0+2k}.$$

To simplify $G_n(z; d_0)$, we have

$$\frac{d}{dz} [z^{x_0+y_0} G_n(z; d_0)] = z^{x_0+y_0-1} \sum_{k=n-1}^{\infty} f_{n,d_0+2k}(d_0) z^{d_0+2k} = z^{x_0+y_0-1} \Phi_n(z; d_0), \quad (\text{C.14})$$

where $\Phi_n(z; d_0) = \sum_{k=n-1}^{\infty} f_{n,d_0+2k}(d_0) z^{d_0+2k}$ is the generating function of the probability distribution of the n -th visit to the origin in a simple random walk starting from d_0 . Let $F_1(z)$ be the generating function of the distribution of the time of the first return to the origin in a simple random walk starting from the origin. The standard renewal argument (see e.g. XI.3.d of [28]) shows that $\Phi_n(z; d_0)$ is given by

$$\Phi_n(z; d_0) = [\Phi_1(z; 1)]^{d_0} [F_1(z)]^{n-1},$$

where $\Phi_1(z; 1)$ and $F_1(z)$ are given by Eqs. (3.6) and (3.14) of [28, Chap. XI], respectively. Therefore,

$$\Phi_n(z; d_0) = z^{-d_0} \left(1 - \sqrt{1 - z^2}\right)^{n+d_0-1}. \quad (\text{C.15})$$

Substituting Eq. (C.15) into Eq. (C.14) and integrating from 0 to 1 yields

$$G_n(1; d_0) = \int_0^1 z^{2\min\{x_0, y_0\}-1} \left(1 - \sqrt{1 - z^2}\right)^{n+d_0-1} dz,$$

where we have used $x_0 + y_0 - d_0 = 2 \min\{x_0, y_0\}$. A change of variable $u = \sqrt{1 - z^2}$ yields

$$G_n(1; d_0) = \int_0^1 u(1 - u^2)^{\min\{x_0, y_0\}-1} (1 - u)^{n+d_0-1} du,$$

which is upper bounded by

$$G_n(1; d_0) \leq \int_0^1 u(1 - u)^{n+d_0-1} du = B(2, n + d_0), \quad (\text{C.16})$$

and lower bounded by

$$G_n(1; d_0) \geq \int_0^1 u(1-u)^{\min\{x_0, y_0\}-1} (1-u)^{n+d_0-1} du = B(2, n + \max\{x_0, y_0\} - 1), \quad (\text{C.17})$$

where we have used $\min\{x_0, y_0\} + d_0 = \max\{x_0, y_0\}$. Applying Eqs. (C.16) and (C.17) to Eq. (C.13) yields

$$\frac{B(2, n + \max\{x_0, y_0\} - 1)}{2^{x_0+y_0-2} B(x_0, y_0)} \leq \mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[N = n] \leq \frac{B(2, n + d_0)}{2^{x_0+y_0-2} B(x_0, y_0)}.$$

Note that $\mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[N = \infty] = \mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[T = \infty] = 0$. Summing over n and using $\sum_{m=n}^{\infty} B(2, m) = n^{-1}$, we obtain

$$\frac{1}{2^{x_0+y_0-2} B(x_0, y_0)} \cdot \frac{1}{n + \max\{x_0, y_0\} - 1} \leq \mathbb{P}_{\text{CA},1}^{(x_0,y_0)}[N \geq n] \leq \frac{1}{2^{x_0+y_0-2} B(x_0, y_0)} \cdot \frac{1}{n + d_0},$$

which immediately yields Eq. (3.6).

C.4 Proof of Theorem 3.4.4

We first prove the following lemma.

Lemma C.4.1. *The probability $\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\tau_1 < \infty]$ of ever having a tie is bounded as follows,*

$$\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\tau_1 < \infty] \leq \begin{cases} 1, & x_0 \leq y_0, \\ \frac{(y_0)_{x_0-y_0}}{(rx_0+y_0)_{x_0-y_0}} \left(1 + \frac{1}{r}\right)^{x_0-y_0}, & x_0 > y_0. \end{cases}$$

Note that $\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\tau_1 < \infty]$ is the exit probability $E(x, y)$ in [6] when $r = 1$.

Proof. The case $x_0 = y_0$ is trivial since $\tau_1 = 0$. When $x_0 < y_0$, Theorem 3.21 of [42] yields $Y_t/X_t \rightarrow 0$ almost surely, from which it follows that $X_t > Y_t$ eventually and hence $\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\tau_1 < \infty] = 1$.

Now assume $x_0 > y_0$. Recall that $A_{1,t}(x_0, y_0)$ is the set of paths starting from (x_0, y_0) that end with the first tie at time t . Note that $A_{1,t}(x_0, y_0)$ is nonempty only if $t = d_0 + 2k$, where $d_0 = x_0 - y_0$ and $k \geq 0$. Let $\pi \in A_{1,t}(x_0, y_0)$ and its state at time j be $\pi_j = (x_j, y_j)$. The probability of the path π is given by

$$\begin{aligned}\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\pi] &= \prod_{j=0}^{t-1} \left(\frac{rx_j}{rx_j + y_j} \right)^{x_{j+1}-x_j} \left(\frac{y_j}{rx_j + y_j} \right)^{y_{j+1}-y_j} \\ &= \frac{r^{x_t-x_0}(x_0)_{x_t-x_0}(y_0)_{y_t-y_0}}{\prod_{j=0}^{t-1} (rx_j + y_j)} \\ &= \frac{r^{x_t-x_0}(x_0)_{x_t-x_0}(y_0)_{y_t-y_0}}{\prod_{j=0}^{t-1} [(r-1)x_j + x_0 + y_0 + j]},\end{aligned}$$

where in the last step we have used $x_j + y_j = x_0 + y_0 + j$. Note that $\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\pi]$ is maximized if the x_j 's are minimized, subject to the constraints that the x_j 's increase monotonically from x_0 to x_t with step size 0 or 1, and that $x_j > y_j$ for all $1 \leq j \leq t-1$, or equivalently $x_j > x_0 + (j - d_0)/2$. This is achieved by the following sequence,

$$x_j^* = \begin{cases} x_0, & j = 0, 1, \dots, d_0 - 1; \\ x_0 + \lfloor (j - d_0)/2 \rfloor + 1, & j = d_0, d_0 + 1, \dots, t - 1; \\ x_t, & j = t. \end{cases}$$

The corresponding path π^* has probability

$$\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\pi^*] = \prod_{j=0}^{d_0-2} \frac{y_0 + j}{rx_0 + y_0 + j} \prod_{x=x_0}^{x_t-1} \frac{rx}{rx + (x-1)} \cdot \frac{x-1}{r(x+1) + (x-1)} \cdot \frac{x_t-1}{rx_t + (x_t-1)},$$

which, after arrangement, yields,

$$\begin{aligned}\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\pi^*] &= \prod_{j=0}^{d_0-1} \frac{y_0 + j}{rx_0 + y_0 + j} \prod_{x=x_0}^{x_t-1} \frac{rx}{(r+1)x + r} \cdot \frac{x}{(r+1)x + (r-1)} \\ &\leq \frac{(y_0)_{d_0}}{(rx_0 + y_0)_{d_0}} \frac{r^{x_t-x_0}}{(r+1)^{2(x_t-x_0)}}\end{aligned}$$

$$\begin{aligned}
&= \frac{(y_0)_{d_0}(r+1)^{d_0}}{(rx_0+y_0)_{d_0}} \left(\frac{r}{r+1}\right)^{x_t-x_0} \left(\frac{1}{r+1}\right)^{y_t-y_0} \\
&= \frac{(y_0)_{d_0}(r+1)^{d_0}}{(rx_0+y_0)_{d_0}} \mathbb{P}_{\text{RW},r}^{(x_0,y_0)}[\pi].
\end{aligned}$$

Thus we have

$$\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\pi] \leq \mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\pi^*] \leq \frac{(y_0)_{d_0}(r+1)^{d_0}}{(rx_0+y_0)_{d_0}} \mathbb{P}_{\text{RW},r}^{(x_0,y_0)}[\pi],$$

and, after summing over $\pi \in A_{1,t}(x_0, y_0)$,

$$\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\tau_1 = t] \leq \frac{(y_0)_{d_0}(r+1)^{d_0}}{(rx_0+y_0)_{d_0}} \mathbb{P}_{\text{RW},r}^{(x_0,y_0)}[\tau_1 = t].$$

Summing over t , we obtain

$$\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\tau_1 < \infty] \leq \frac{(y_0)_{d_0}(r+1)^{d_0}}{(rx_0+y_0)_{d_0}} \mathbb{P}_{\text{RW},r}^{(x_0,y_0)}[\tau_1 < \infty].$$

By Eq. (3.9) and XI.3.d of [28], $\mathbb{P}_{\text{RW},r}^{(x_0,y_0)}[\tau_1 < \infty] = r^{-d_0}$, from which the desired conclusion follows. \square

Corollary C.4.2. *The probability of having at least one more tie starting from a tie state (x, x) is bounded by*

$$\mathbb{P}_{\text{CA},r}^{(x,x)}[\tau_2 < \infty] \leq \frac{2}{r+1}.$$

Proof. By considering the one-step transition from (x, x) into $(x, x+1)$ or $(x+1, x)$, we obtain

$$\begin{aligned}
\mathbb{P}_{\text{CA},r}^{(x,x)}[\tau_2 < \infty] &= \frac{r}{r+1} \mathbb{P}_{\text{CA},r}^{(x+1,x)}[\tau_1 < \infty] + \frac{1}{r+1} \mathbb{P}_{\text{CA},r}^{(x,x+1)}[\tau_1 < \infty] \\
&\leq \frac{x}{(r+1)x+r} + \frac{1}{r+1} \leq \frac{2}{r+1},
\end{aligned}$$

where the first inequality follows from Lemma C.4.1. \square

Now we prove Theorem 3.4.4.

Proof of Theorem 3.4.4. Let Z_n be the common value of X_t and Y_t at $t = \tau_n$, i.e., $Z_n = X_{\tau_n}$. Conditioned on $\tau_n < \infty$ and $Z_n = z$, the probability of $\tau_{n+1} < \infty$ is just the probability of having a tie after leaving (z, z) . Thus

$$\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\tau_{n+1} < \infty \mid \tau_n < \infty, Z_n = z] = \mathbb{P}_{\text{CA},r}^{(z,z)}[\tau_2 < \infty] \leq \frac{2}{r+1},$$

by Corollary C.4.2. Removal of the conditioning yields

$$\mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\tau_{n+1} < \infty \mid \tau_n < \infty] \leq \frac{2}{r+1}.$$

It follows that

$$\begin{aligned} \mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[N \geq n] &= \mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\tau_n < \infty] \\ &= \mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\tau_1 < \infty] \prod_{i=1}^{n-1} \mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\tau_{i+1} < \infty \mid \tau_i < \infty] \\ &\leq \mathbb{P}_{\text{CA},r}^{(x_0,y_0)}[\tau_1 < \infty] \left(\frac{2}{r+1}\right)^{n-1}. \end{aligned}$$

An application of Lemma C.4.1 completes the proof. \square

APPENDIX D

ADDITIONAL PROOFS FOR CHAPTER 4

This appendix provides proofs of the theorems in Section 4.3.1.

D.1 Proof of Theorem 4.3.1

Let $p_i = \Lambda^{-1} \sum_{j=1}^i \lambda_j$. Note that $\Lambda^{-1} \lambda_{i+1,i} = 1 - p_i$ and $\Lambda^{-1} \lambda_{i-1,i} = p_{i-1}$. Also note that $n_{i,i+1} = i$ and $n_{i,i-1} = n - i + 1$. Thus the optimal social cost in (4.8) can be rewritten as

$$\begin{aligned} C^* &= \frac{1}{n-1} \sum_{i=1}^n \frac{1}{b_i} \left[\sqrt{i(1-p_i)} + \sqrt{(n-i+1)p_{i-1}} \right]^2 \\ &\geq \frac{1}{(n-1)b_{\max}} \sum_{i=1}^n \left[\sqrt{i(1-p_i)} + \sqrt{(n-i+1)p_{i-1}} \right]^2. \end{aligned}$$

The inequality $(x+y)^2 \geq x^2 + y^2$ for $xy \geq 0$ then yields

$$C^* \geq \frac{C^o}{b_{\max}}, \tag{D.1}$$

where

$$C^o = \frac{1}{n-1} \sum_{i=1}^n [i(1-p_i) + (n-i+1)p_{i-1}]. \tag{D.2}$$

Let $i^* = \min\{i : p_i \geq 1/2\}$. Since p_i is increasing in i , we have $p_i < 1/2$ for all $i < i^*$ and $p_i \geq 1/2$ for all $i \geq i^*$. Therefore,

$$C^o \geq \frac{1}{n-1} \left[\sum_{i=1}^{i^*-1} \frac{1}{2}i + \sum_{i=i^*+1}^n \frac{1}{2}(n-i+1) \right]$$

$$= \frac{1}{2(n-1)} \left[\left(i^* - \frac{n+1}{2} \right)^2 + \frac{n^2 - 1}{4} \right] \geq \frac{n+1}{8}, \quad (\text{D.3})$$

which, combined with (D.1), yields the lower bound in (4.9).

The cost under selfish allocation in (4.6) can be upper bounded as follows,

$$\begin{aligned} \hat{C} &= \frac{1}{n-1} \sum_{i=1}^n \frac{1}{b_i} \left[i\sqrt{1-p_i} + (n-i+1)\sqrt{p_{i-1}} \right] \left(\sqrt{1-p_i} + \sqrt{p_{i-1}} \right) \\ &\leq \frac{1}{(n-1)b_{\min}} \sum_{i=1}^n \left[i\sqrt{1-p_i} + (n-i+1)\sqrt{p_{i-1}} \right] \left(\sqrt{1-p_i} + \sqrt{p_{i-1}} \right) \\ &= \frac{1}{(n-1)b_{\min}} \sum_{i=1}^n \left[i(1-p_i) + (n-i+1)p_{i-1} + (n+1)\sqrt{p_{i-1}(1-p_i)} \right] \\ &= \frac{C^o + C^d}{b_{\min}}, \end{aligned} \quad (\text{D.4})$$

where C^o is as in (D.2) and

$$C^d = \frac{n+1}{n-1} \sum_{i=2}^{n-1} \sqrt{p_{i-1}(1-p_i)}. \quad (\text{D.5})$$

Since $\sqrt{p_{i-1}(1-p_i)} \leq 2^{-1}(p_{i-1} + 1 - p_i) \leq 2^{-1}$,

$$C^d \leq \frac{(n+1)(n-2)}{2(n-1)}. \quad (\text{D.6})$$

For an upper bound on C^o , we again use $p_i < 1/2$ for $i < i^*$ and $p_i \geq 1/2$ for $i \geq i^*$.

Thus

$$\begin{aligned} C^o &= \frac{1}{n-1} \sum_{i=1}^n [p_{i-1}(n-i+1) + i(1-p_i)] \\ &\leq \frac{1}{n-1} \left[\sum_{i=1}^{i^*} \frac{1}{2}(n-i+1) + \sum_{i=i^*}^n \frac{1}{2}i \right] \\ &= \frac{1}{2(n-1)} \left[\frac{5n^2 + 6n + 1}{4} - \left(i^* - \frac{n+1}{2} \right)^2 \right] \end{aligned}$$

$$\leq \frac{(n+1)(5n+1)}{8(n-1)}, \quad (\text{D.7})$$

which, together with (D.4) and (D.6), yields the upper bound in (4.9).

By (D.1) and (D.4), we have

$$\text{PoS} = \frac{\widehat{C}}{C^*} \leq \frac{b_{\max}}{b_{\min}} \left(1 + \frac{C^d}{C^o} \right).$$

Thus (4.10) follows from (D.3) and (D.6).

D.2 Proof of Theorem 4.3.2

Consider the case of homogeneous content creation rates with $\lambda_i = 1$. Consider a heterogeneous set of budgets of allocation where $b_i = b_1 > b_2$, for all $i \neq 2$. The optimal social cost, under centralized allocation, is given by

$$C^* = \frac{C^o}{b_1} + \frac{1}{b_1} \left(\frac{b_1}{b_2} - 1 \right) \frac{1}{n-1} \left[\sqrt{2 \left(1 - \frac{2}{n} \right)} + \sqrt{1 - \frac{1}{n}} \right]^2,$$

where C^o is as in (D.2). The social cost under selfish allocation is

$$\widehat{C} = \frac{C^o + C^d}{b_1} + \frac{1}{b_1} \left(\frac{b_1}{b_2} - 1 \right) \frac{1}{n-1} \left(\frac{1}{\sqrt{n}} + \sqrt{1 - \frac{2}{n}} \right) \left[\left(\sqrt{n} - \frac{1}{\sqrt{n}} \right) + 2\sqrt{1 - \frac{2}{n}} \right],$$

where C^d is as in (D.5). By (D.3), (D.7) and (D.6), $C^o = \Theta(n)$ and $C^d = O(n)$. If $b_2/b_1 = \Omega(n^2)$, then $C^* = \Theta(b_2^{-1}n^{-1})$ and $\widehat{C} = \Theta(b_2^{-1}n^{-1/2})$. Thus $\text{PoS} = \widehat{C}/C^* = \Theta(\sqrt{n})$, which grows unbounded with the network size.

D.3 Proof of Theorem 4.3.3

Since b changes C^* and \widehat{C} only by a multiplicative factor, it suffices to prove the theorem for $b = 1$. Let i be the hub of the i -th star, and L_i the leaf nodes connected

to it. Then $n_{il} = n - 1$ and $n_{li} = 1$ for any $l \in L_i$. For adjacent hubs, $n_{i-1,i} = (i-1)p$ and $n_{i+1,i} = (k-i)p$. Then the optimal social cost is given by

$$C^* = \frac{1}{n(n-1)} \left(\sum_{i=1}^k \sum_{l \in L_i} (n-1) + S_1 \right) = 1 - \frac{k}{n} + \frac{S_1}{n(n-1)},$$

where

$$S_1 = \sum_{i=1}^k \left[\sum_{l \in L_i} \sqrt{n-1} + \sum_{j=i \pm 1} \sqrt{(n-n_{ji})n_{ji}} \right]^2.$$

Using the inequalities $\sum_{i=1}^3 a_i^2 \leq (\sum_{i=1}^3 a_i)^2 \leq 3 \sum_{i=1}^3 a_i^2$, we obtain $S_1 = \Theta(S_2)$,

where

$$\begin{aligned} S_2 &= k(p-1)^2(n-1) + \sum_{i=1}^k \sum_{j=i-1}^i p^2 j(k-j) \\ &= (n-k)(p-1)(n-1) + \frac{1}{3}n(nk-p) = \Theta(n^2 \max\{p, k\}). \end{aligned}$$

Thus $C^* = \Theta(\max\{p, k\}) = \Omega(\sqrt{n})$.

The cost under selfish allocation can be written as follows,

$$\widehat{C} = \frac{1}{n(n-1)} \left(\sum_{i=1}^k \sum_{l \in L_i} (n-1) + S_3 \right) = 1 - \frac{k}{n} + \frac{S_3}{n(n-1)},$$

where

$$S_3 = \sum_{i=1}^k \left(\sum_{l \in L_i} 1 + \sum_{j=i \pm 1} \sqrt{n_{ji}} \right) \left[\sum_{l \in L_i} (n-1) + \sum_{j=i \pm 1} (n-n_{ji})\sqrt{n_{ji}} \right].$$

Using the facts that $|L_i| = p-1$, $n_{ji} \leq n$ and the inequality $(n-x)\sqrt{x} \leq \frac{2}{3\sqrt{3}}n^{3/2}$ for $x \in [0, n]$, we obtain

$$S_3 \leq \sum_{i=1}^k (p+2\sqrt{n}) \left(pn + \frac{4}{3\sqrt{3}}n^{3/2} \right)$$

$$= O \left((n + k\sqrt{n})(pn + n^{3/2}) \right) = O \left(n^2 \max\{k, p\} \right).$$

Thus $\widehat{C} = O(\max\{p, k\})$. Since $\widehat{C} \geq C^* = \Theta(\max\{p, k\})$, we have $\widehat{C} = \Theta(\max\{p, k\})$ and PoS = $\Theta(1)$.

D.4 Proof of Theorem 4.3.4

As in the proof of Theorem 4.3.3, it suffices to consider the case $b = 1$. Note that $\lambda_{ji} = n_{ji}$. The optimal social cost (4.8) and cost under selfish allocation (4.6) can now be written as follows,

$$C^* = \frac{1}{n(n-1)} \sum_{i \in V} \left(\sum_{j \in N(i)} \sqrt{(n - n_{ji})n_{ji}} \right)^2, \quad (\text{D.8})$$

and

$$\widehat{C} = \frac{1}{n(n-1)} \sum_{i \in V} \left[\sum_{j \in N(i)} (n - n_{ji}) \sqrt{n_{ji}} \right] \left(\sum_{j \in N(i)} \sqrt{n_{ji}} \right). \quad (\text{D.9})$$

Let the depth of the tree be h . Then $n = \frac{k^{h+1}-1}{k-1}$. Label the nodes in such a way that for $i = 1, 2, \dots, n$, $p(i) = \lceil (i-1)/k \rceil$ is the parent of i , $c_j(i) = k(i-1)+j+1$ is the j -th child of i , for $j = 1, 2, \dots, k$. Then $n_{p(i),i} = \frac{k^{h+1}-k^{h-h(i)+1}}{k-1}$ and $n_{c_j(i),i} = \frac{k^{h-h(i)-1}}{k-1}$, where $h(i) = \lfloor \log_k(ki - i) \rfloor$ is the depth of node i .

By (D.8), the optimal social cost is

$$C^* = \frac{1}{n(n-1)} \sum_{i=1}^n \left[\sqrt{(n - n_{p(i),i})n_{p(i),i}} + \sum_{j=1}^k \sqrt{(n - n_{c_j(i),i})n_{c_j(i),i}} \right]^2,$$

which is bounded by

$$\frac{1}{n(n-1)}(S_1 + k^2 S_2) \leq C^* \leq \frac{2}{n(n-1)}(S_1 + k^2 S_2),$$

where

$$S_1 = \sum_{i=1}^n (n - n_{p(i),i})n_{p(i),i},$$

and

$$S_2 = \sum_{i=1}^n (n - n_{c_1(i),i}) n_{c_1(i),i}.$$

Using the facts that $n = \frac{k^{h+1}-1}{k-1}$, $n_{p(i),i} = \frac{k^{h+1}-k^{h-h(i)+1}}{k-1}$ and $n_{c_1(i),i} = \frac{k^{h-h(i)}-1}{k-1}$, we obtain

$$\begin{aligned} S_1 &= \frac{1}{(k-1)^2} \sum_{i=1}^n (k^{h-h(i)+1} - 1)(k^{h+1} - k^{h-h(i)+1}) \\ &= \frac{1}{(k-1)^2} \sum_{h'=0}^h k^{h'} (k^{h-h'+1} - 1)(k^{h+1} - k^{h-h'+1}) \\ &= \frac{hk - h - 2}{(k-1)^3} k^{2h+2} + \frac{hk - h + 2k}{(k-1)^3} k^{h+1} \\ &= \Theta(hk^{2h}) = \Theta(n^2 \log_k n), \end{aligned}$$

and

$$\begin{aligned} S_2 &= \frac{1}{(k-1)^2} \sum_{i=1}^n (k^{h+1} - k^{h-h(i)}) (k^{h-h(i)} - 1) \\ &= \frac{1}{(k-1)^2} \sum_{h'=0}^h k^{h'} (k^{h+1} - k^{h-h'}) (k^{h-h'} - 1) = \frac{1}{k} S_1. \end{aligned}$$

Therefore,

$$\frac{1+k}{n(n-1)} S_1 \leq C^* \leq \frac{2(1+k)}{n(n-1)} S_1,$$

and hence $C^* = \Theta(k \log_k n)$.

By (D.9), the social cost under selfish allocation is

$$\begin{aligned} \widehat{C} &= \frac{1}{n(n-1)} \sum_{i=1}^n [(n - n_{p(i),i}) \sqrt{n_{p(i),i}} + k(n - n_{c_1(i),i}) \sqrt{n_{c_1(i),i}}] (\sqrt{n_{p(i),i}} + 2\sqrt{n_{c_1(i),i}}) \\ &= \frac{1}{n(n-1)} (S_1 + k^2 S_2 + k S_3), \end{aligned}$$

where

$$S_3 = \sum_{i=1}^n (2n - n_{c_1(i),i} - n_{p(i),i}) \sqrt{n_{c_1(i),i} n_{p(i),i}}.$$

Note that $2n \geq 2n - n_{c_1(i),i} - n_{p(i),i} \geq n$, and $n \geq n_{p(i),i} \geq n/2$ for $i \neq 1$, so

$$S_3 = \Theta\left(n^{3/2} \sum_{i=2}^n \sqrt{n_{c_1(i),i}}\right).$$

We also have

$$\sum_{i=2}^n \sqrt{n_{c_1(i),i}} = \frac{1}{\sqrt{k-1}} \sum_{h'=1}^{h-1} k^{h'} \sqrt{k^{h-h'} - 1} = \Theta(n/k),$$

yielding $S_3 = \Theta(kn^2 \log_k n + n^{5/2})$ and hence $\hat{C} = \Theta(S_3/n^2) = \Theta(k \log_k n + \sqrt{n})$.

Taking the ratio \hat{C}/C^* , the result follows.

BIBLIOGRAPHY

- [1] <http://dumps.wikimedia.org/>.
- [2] Klout. <http://klout.com>.
- [3] Albert, Réka, Jeong, Hawoong, and Barabási, Albert-László. Internet: Diameter of the world-wide web. *Nature* 401, 6749 (1999), 130–131.
- [4] Amini, Hamed, Draief, Moez, and Lelarge, Marc. Flooding in weighted random graphs. In *Proc. SIAM ANALCO* (2011), pp. 1–15.
- [5] Anstee, Richard P. Properties of a class of (0, 1)-matrices covering a given matrix. *Can. J. Math* 34, 2 (1982), 438–453.
- [6] Antal, Tibor, Ben-Naim, Eli, and Krapivsky, Pavel L. First-passage properties of the Pólya urn process. *Journal of Statistical Mechanics: Theory and Experiment* 2010, 07 (2010), P07009.
- [7] Arthur, W Brian. *Increasing Returns and Path Dependence in the Economy*. U. Michigan Press, 1994.
- [8] Backstrom, Lars, Bakshy, Eytan, Kleinberg, Jon, Lento, Thomas M., and Rosen, Itamar. Center of attention: How facebook users allocate attention across friends. In *Proc. 5th International Conference on Weblogs and Social Media* (2011).
- [9] Barabási, Albert-László. Network science: Luck or reason. *Nature* 489, 7417 (2012), 507–508.
- [10] Barabási, Albert-László, and Albert, Réka. Emergence of scaling in random networks. *Science* 286, 5439 (1999), 509–512.
- [11] Berge, Claude. *Graphs and hypergraphs*, vol. 7. North-Holland Publishing Company Amsterdam, 1973.
- [12] Bianconi, Ginestra, and Barabási, Albert-László. Competition and multiscaling in evolving networks. *Europh. Let.* 54, 4 (2001), 436–442.
- [13] Blondel, Vincent D., Guillaume, Jean-Loup, Hendrickx, Julien M., de Kerchove, Cristobald, and Lambiotte, Renaud. Local leaders in random networks. *Phys. Rev. E* 77 (Mar 2008), 036114.

- [14] Borghol, Youmna, Mitra, Siddharth, Ardon, Sébastien, Carlsson, Niklas, Eager, Derek, and Mahanti, Anirban. Characterizing and modelling popularity of user-generated videos. *Performance Evaluation* 68, 11 (2011), 1037–1055.
- [15] Bubeck, Sébastien, Mossel, Elchanan, and Rácz, Miklós Z. On the influence of the seed graph in the preferential attachment model. *Network Science and Engineering, IEEE Transactions on PP*, 99 (2015), 1–1.
- [16] Busch, Arthur H, Ferrara, Michael J, Hartke, Stephen G, Jacobson, Michael S, Kaul, Hemanshu, and West, Douglas B. Packing of graphic n -tuples. *Journal of Graph Theory* 70, 1 (2012), 29–39.
- [17] Censor-Hillel, Keren, and Shachnai, Hadas. Fast information spreading in graphs with large weak conductance. In *Proceedings of ACM-SIAM SODA* (2011).
- [18] Cha, Meeyoung, Mislove, Alan, and Gummadi, Krishna P. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th International Conference on World Wide Web* (2009), pp. 721–730.
- [19] Chen, Wai-Kai. On the realization of a (p, s)-digraph with prescribed degrees. *Journal of the Franklin Institute* 281, 5 (1966), 406–422.
- [20] Chierichetti, Flavio, Lattanzi, Silvio, and Panconesi, Alessandro. Rumor spreading in social networks. *Automata, Languages and Programming* (2009), 375–386.
- [21] Denrell, Jerker, and Liu, Chengwei. Top performers are not the most impressive when extreme performance indicates unreliability. *Proceedings of the National Academy of Sciences* 109, 24 (2012), 9331–9336.
- [22] DiPrete, Thomas A, and Eirich, Gregory M. Cumulative advantage as a mechanism for inequality: A review of theoretical and empirical developments. *Annual review of sociology* (2006), 271–297.
- [23] Drinea, Eleni, Frieze, Alan, and Mitzenmacher, Michael. Balls and bins models with feedback. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)* (2002), pp. 308–315.
- [24] Dürr, Christoph, Guiñez, Flavio, and Matamala, Martín. Reconstructing 3-colored grids from horizontal and vertical projections is NP-hard. In *Algorithms-ESA 2009*, vol. 5757 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2009, pp. 776–787.
- [25] Eggenberger, Florian, and Pólya, George. Über die statistik verketteter vorgänge. *Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik* 3, 4 (1923), 279–289.
- [26] Eisenberg, Eli, and Levanon, Erez Y. Preferential attachment in the protein network evolution. *Phy. Rev. Let.* 91, 13 (2003), 138701.

[27] Erdős, Paul, and Gallai, Tibor. Gráfok előírt fokszámú pontokkal (graphs with prescribed degrees). *Matematikai Lapok* 11 (1960), 264–274.

[28] Feller, Willliam. *An introduction to probability theory and its applications*, 3rd ed., vol. 1. John Wiley & Sons, 1968.

[29] Figueiredo, Flavio, Benevenuto, Fabrício, and Almeida, Jussara M. The tube over time: characterizing popularity growth of youtube videos. In *ACM international conference on Web search and data mining* (2011), pp. 745–754.

[30] Fulkerson, Delbert Ray, et al. Zero-one matrices with zero trace. *Pacific J. Math* 10, 3 (1960), 831–836.

[31] Garlaschelli, Diego, and Loffredo, Maria I. Patterns of link reciprocity in directed networks. *Phys. Rev. Lett.* 93 (Dec 2004), 268701.

[32] Gaudeul, Alexia, and Peroni, Chiara. Reciprocal attention and norm of reciprocity in blogging networks. Tech. rep., Jena Economic Research Papers, 2010.

[33] Godrèche, Claude, Grandclaude, Hélène, and Luck, Jean-Marc. Statistics of leaders and lead changes in growing networks. *Journal of Statistical Mechanics: Theory and Experiment* 2010, 02 (2010), P02001.

[34] Godrèche, Claude, and Luck, Jean-Marc. On leaders and condensates in a growing network. *Journal of Statistical Mechanics: Theory and Experiment* 2010, 07 (2010), P07031.

[35] Gonzalez, Roberto, Cuevas, Ruben, Motamedi, Reza, Rejaie, Reza, and Cuevas, Angel. Google+ or Google-?: Dissecting the evolution of the new OSN in its first year. In *Proceedings of the 22Nd International Conference on World Wide Web* (2013), WWW '13, pp. 483–494.

[36] Hakimi, S Louis. On realizability of a set of integers as degrees of the vertices of a linear graph. i. *Journal of the Society for Industrial & Applied Mathematics* 10, 3 (1962), 496–506.

[37] Havel, Václav. A remark on the existence of finite graphs (in czech). *Casopis Pest. Mat* 80, 477-480 (1955), 1253.

[38] Hughes, Barry D. *Random Walks and Random Environments*, vol. 1: Random Walks. Oxford University Press, 1995.

[39] Ioannidis, Stratis, Chaintreau, Augustin, and Massoulié, Laurent. Optimal and scalable distribution of content updates over a mobile social network. In *IEEE INFOCOM'09* (April 2009), pp. 1422 –1430.

[40] Iványi, Antal, and Lucz, Loránd. Erdos-Gallai test in linear time. *Combinatorica* (2011).

- [41] Janson, Svante. One, two and three times $\log n/n$ for paths in a complete graph with random weights. *Comb. Probab. Comput.* 8, 4 (July 1999), 347–361.
- [42] Janson, Svante. Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stochastic Processes and their Applications* 110, 2 (2004), 177–245.
- [43] Janson, Svante. Limit theorems for triangular urn schemes. *Prob. Theory Related Fields* 134 (2005), 417–452.
- [44] Java, Akshay, Song, Xiaodan, Finin, Tim, and Tseng, Belle. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis* (2007), pp. 56–65.
- [45] Johnson, Norman L, Kemp, Adrienne W, and Kotz, Samuel. *Univariate Discrete Distributions*, 2nd ed. John Wiley & Sons, 1992.
- [46] Kampen, NG van. Stochastic processes in physics and chemistry. *North-Holland* 1 (1981).
- [47] Karp, Richard, Schindelhauer, Christian, Shenker, Scott, and Vocking, Berthold. Randomized rumor spreading. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on* (2000), IEEE, pp. 565–574.
- [48] Kleitman, Daniel J, and Wang, Da-Lun. Algorithms for constructing graphs and digraphs with given valences and factors. *Discrete Mathematics* 6, 1 (1973), 79–88.
- [49] Kossinets, Gueorgi, Kleinberg, Jon, and Watts, Duncan. The structure of information pathways in a social communication network. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (2008), ACM, pp. 435–443.
- [50] Krapivsky, Pavel L, and Redner, Sidney. Statistics of changes in lead node in connectivity-driven networks. *Physical review letters* 89, 25 (2002), 258703.
- [51] Kwak, Haewoon, Lee, Changhyun, Park, Hosung, and Moon, Sue. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web* (2010), pp. 591–600.
- [52] Leskovec, Jure, and Krevl, Andrej. SNAP Datasets: Stanford large network dataset collection. <http://snap.stanford.edu/data>, June 2014.
- [53] Magno, Gabriel, Comarella, Giovanni, Saez-Trumper, Diego, Cha, Meeyoung, and Almeida, Virgilio. New kid on the block: Exploring the Google+ social graph. In *Proceedings of the 2012 ACM Conference on Internet Measurement Conference* (2012), pp. 159–170.

- [54] Mahmoud, Hosam. *Pólya urn models*. CRC Press, 2008.
- [55] Merton, Robert K. The matthew effect in science. *Science* 159 (1968), 56–63.
- [56] Mislove, Alan, Marcon, Massimiliano, Gummadi, Krishna P., Druschel, Peter, and Bhattacharjee, Bobby. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement* (2007), pp. 29–42.
- [57] Newman, Mark E. J., Forrest, Stephanie, and Balthrop, Justin. Email networks and the spread of computer viruses. *Phys. Rev. E* 66 (Sep 2002), 035101.
- [58] Nisan, Noam, Roughgarden, Tim, Tardos, Éva, and Vazirani, Vijay V., Eds. *Algorithmic Game Theory*. Cambridge Univ Press, 2007.
- [59] Oliveira, Roberto. Balls-in-bins processes with feedback and brownian motion. *Journal of Combinatorics, Probability and Computing* 17, 1 (2008).
- [60] Olver, Frank W.J. *NIST handbook of mathematical functions*. Cambridge University Press, 2010.
- [61] Pemantle, Robin. A survey of random processes with reinforcement. *Probability Surveys* 4, 1-79 (2007), 25.
- [62] Perc, Matjaž. The matthew effect in empirical data. *Journal of The Royal Society Interface* 11, 98 (2014).
- [63] Petersen, Alexander M., Jung, Woo-Sung, Yang, Jae-Suk, and Stanley, H. Eugene. Quantitative and empirical demonstration of the matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences* 108, 1 (2011), 18–23.
- [64] Pittel, Boris. On spreading a rumor. *SIAM J. Appl. Math.* 47, 1 (Mar. 1987), 213–223.
- [65] Price, Derek J. de Solla. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* 27, 5 (1976), 292–306.
- [66] Qu, Yu, Zheng, Qinghua, Liu, Ting, Li, Jian, and Guan, Xiaohong. In-depth measurement and analysis on densification power law of software execution. In *Proceedings of the 5th International Workshop on Emerging Trends in Software Metrics* (New York, NY, USA, 2014), WETSoM 2014, ACM, pp. 55–58.
- [67] Salganik, Matthew J., Dodds, Peter Sheridan, and Watts, Duncan J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311, 5762 (2006), 854–856.

[68] Teixeira, Miguel Cacho, Monteiro, Pedro Tiago, Guerreiro, Joana Fernandes, Gonçalves, Joana Pinho, Mira, Nuno Pereira, dos Santos, Sandra Costa, Cabrito, Tânia Rodrigues, Palma, Margarida, Costa, Catarina, Francisco, Alexandre Paulo, et al. The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in *Saccharomyces cerevisiae*. *Nucleic Acids Research* **43**, D1 (2014), D161–D166.

[69] van de Rijt, Arnout, Kang, Soong Moon, Restivo, Michael, and Patil, Akshay. Field experiments of success-breeds-success dynamics. *Proceedings of the National Academy of Sciences* **111**, 19 (2014), 6934–6939.

[70] Vinayagam, Arunachalam, Stelzl, Ulrich, Foulle, Raphaele, Plassmann, Stephanie, Zenkner, Martina, Timm, Jan, Assmus, Heike E, Andrade-Navarro, Miguel A, and Wanker, Erich E. A directed protein interaction network for investigating intracellular signal transduction. *Science signaling* **4**, 189 (2011), rs8.

[71] Wallstrom, Timothy C. The equalization probability of the Pólya urn. *The American Mathematical Monthly* **119**, 6 (2012), 516–518.

[72] Wang, Dashun, Song, Chaoming, and Barabási, Albert-László. Quantifying long-term scientific impact. *Science* **342**, 6154 (2013), 127–132.

[73] Watts, Duncan J, and Strogatz, Steven H. Collective dynamics of ‘small-world’ networks. *Nature* **393**, 6684 (1998), 440–442.

[74] White, John G, Southgate, Eileen, Thomson, J Nichol, and Brenner, Sydney. The structure of the nervous system of the nematode *caenorhabditis elegans*. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **314**, 1165 (1986), 1–340.

[75] Yan, Koon-Kiu, Fang, Gang, Bhardwaj, Nitin, Alexander, Roger P, and Gerstein, Mark. Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. *Proceedings of the National Academy of Sciences* **107**, 20 (2010), 9186–9191.

[76] Yilmaz, Alper, Mejia-Guerra, Maria Katherine, Kurz, Kyle, Liang, Xiaoyu, Welch, Lonnie, and Grotewold, Erich. AGRIS: Arabidopsis Gene Regulatory Information Server, an update. *Nucleic Acids Research* **39**, suppl 1 (2011), D1118–D1122.

[77] Ying, Annie T. T. Mining challenge 2015: Comparing and combining different information sources on the stack overflow data set. In *The 12th Working Conference on Mining Software Repositories* (2015).

[78] Zakharov, Pavel. Structure of LiveJournal social network. In *SPIE Fourth International Symposium on Fluctuations and Noise* (2007), International Society for Optics and Photonics, pp. 660109–660109.

- [79] Zamora-López, Gorka, Zlatić, Vinko, Zhou, Changsong, Štefančić, Hrvoje, and Kurths, Jürgen. Reciprocity of networks with degree correlations and arbitrary degree sequences. *Phys. Rev. E* **77** (Jan 2008), 016106.
- [80] Zingg, Brian, Hintiryan, Hour, Gou, Lin, Song, Monica Y, Bay, Maxwell, Bienkowski, Michael S, Foster, Nicholas N, Yamashita, Seita, Bowman, Ian, Toga, Arthur W, et al. Neural networks of the mouse neocortex. *Cell* **156**, 5 (2014), 1096–1111.
- [81] Zlatić, Vinko, Božičević, Miran, Štefančić, Hrvoje, and Domazet, Mladen. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Phys. Rev. E* **74** (Jul 2006), 016115.
- [82] Zlatić, Vinko, and Štefančić, Hrvoje. Model of wikipedia growth based on information exchange via reciprocal arcs. *Europhysics Letters* **93**, 5 (2011), 58005.